

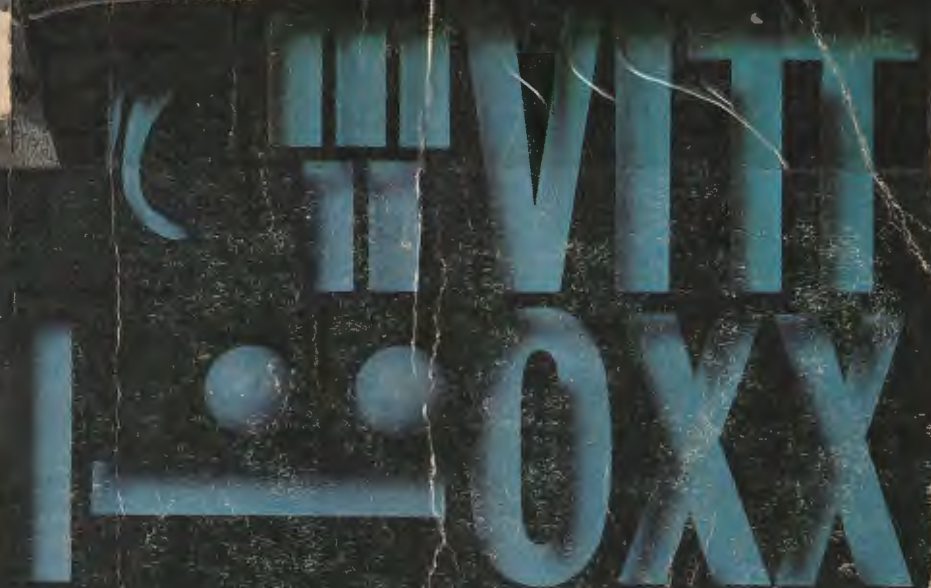
ed. D. Aleksandrov, A.N. Kolmogorov, M.A. Laurentiev y otros

La matemática:

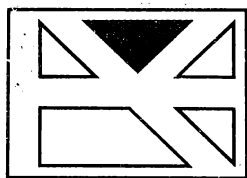
su contenido, métodos y significado

Alianza Universidad

3



LA MATEMATICA: SU CONTENIDO, METODOS Y SIGNIFICADO ofrece una panorámica general difícilmente mejorable de ese vasto campo temático. Publicada originalmente en la Unión Soviética, su traducción al inglés —bajo el patrocinio de la «American Mathematical Society»— ha sido saludada con grandes elogios por las revistas especializadas: «una principalísima contribución a la cultura científica de nuestro siglo» (Harry Schwartz en «The New York Times Book Review»); «la mejor respuesta dada en forma de libro para la enseñanza de las matemáticas tanto en nivel elemental como en el avanzado» (Morris Kline); «una formidable y fabulosa obra maestra» («Social Science»). La obra se mueve en dos planos: el histórico, que examina el desarrollo de la disciplina desde los orígenes hasta nuestros días; y el expositivo, que logra una organización a gran escala de la matemática moderna. A lo largo de los tres volúmenes (subdivididos en seis partes y diecinueve capítulos), los autores —entre los que figuran P. S. ALEKSANDROV, I. M. GELFAND, A. N. KOLMOGOROV y M. A. LAURENTIEV —«los más destacados nombres de las matemáticas rusas se dan cita en el libro», ha señalado «Science Fortnightly»— subrayan la relación directísima, eclipsada en ocasiones por su alto nivel de abstracción, que las matemáticas guardan con el mundo material. La obra ofrece, a la vez, grandes virtualidades didácticas; como ha señalado Roger S. Pinkham, «su elevada calidad expositiva y su insistencia en los conceptos más que en las demostraciones contribuirán decisivamente a elevar el nivel matemático de todos los que la estudien».



Alianza Editorial

La matemática:
su contenido,
métodos y
significado

3.

Alianza Universidad

A. D. Aleksandrov
A. N. Kolmogorov,
M. A. Laurentiev
y otros

La matemática:
su contenido,
métodos y
significado
3

Versión española de
Andrés Ruiz Merino

Alianza
Editorial

Título de la versión inglesa:

Mathematics: Its Content, Methods, and Meaning

© The Massachusetts Institute of Technology

© Ed. cast.: Alianza Editorial, S. A., Madrid, 1973

Calle Milán, 38; ☎ 200 00 45

ISBN 84-206-2993-6 (Obra completa)

ISBN 84-206-2070-X (Tomo III)

Depósito legal: M. 12.740-1974

Impreso en Ediciones Castilla, S. A., Maestro Alonso, 21. Madrid

Printed in Spain

INDICE

Prólogo a la edición rusa	9
<i>Parte 5</i>	
14. TEORIA DE FUNCIONES DE UNA VARIABLE REAL	
S. B. Steckin	
§1. Introducción	15
§2. Conjuntos	17
§3. Números reales	25
§4. Conjuntos de puntos	32
§5. Medida de conjuntos	41
§6. La integral de Lebesgue	47
15. ALGEBRA LINEAL	
D. K. Fadiev	
§1. El alcance del álgebra lineal y sus métodos	55
§2. Espacios lineales	68
§3. Sistemas de ecuaciones lineales	83
§4. Transformaciones lineales	98
§5. Formas cuadráticas	109
§6. Funciones de matrices y algunas de sus aplicaciones	117
16. GEOMETRIAS NO EUCLIDIANAS	
A. D. Aleksandrov	
§1. Historia del postulado de Euclides	124
§2. La solución de Lobachevski	127
§3. La geometría de Lobachevski	132

§4. El significado real de la geometría de Lobachevski.....	142
§5. Los axiomas de la geometría; su verificación en el presente caso.	151
§6. Separación de las teorías geométricas independientes de la geometría euclidiana.....	159
§7. Espacios multidimensionales.....	168
§8. Generalización del alcance de la geometría.....	185
§9. Geometría riemanniana.....	199
§10. La geometría abstracta y el espacio real.....	215

Parte 6

17. TOPOLOGIA

P. S. Aleksandrov

§1. El objeto de la topología.....	231
§2. Superficies.....	235
§3. Variedades.....	241
§4. El método combinatorio.....	243
§5. Campos vectoriales.....	252
§6. El desarrollo de la topología.....	258
§7. Espacios topológicos y métricos.....	262

18. ANALISIS FUNCIONAL

I. M. Gelfand

§1. El espacio n -dimensional.....	268
§2. Espacio de Hilbert (Espacio infinito-dimensional).....	273
§3. Desarrollo mediante sistemas ortogonales de funciones.....	279
§4. Ecuaciones integrales.....	288
§5. Operadores lineales y desarrollos ulteriores del análisis funcional.....	296

19. GRUPOS Y OTROS SISTEMAS ALGEBRAICOS

A. I. Malsev

§1. Introducción.....	309
§2. Simetrías y transformaciones.....	311
§3. Grupos de transformaciones.....	321
§4. Grupos de Fedorov (Grupos cristalográficos).....	334
§5. Grupos de Galois.....	343
§6. Conceptos fundamentales de la teoría general de grupos...	347
§7. Grupos continuos.....	356
§8. Grupos fundamentales.....	359
§9. Representaciones y caracteres de grupos.....	367
§10. La teoría general de grupos.....	373
§11. Números hipercomplejos.....	374
§12. Algebras asociativas.....	385
§13. Algebras de Lie.....	395
§14. Anillos.....	398
§15. Retículos.....	404
§16. Otros sistemas algebraicos.....	406

PROLOGO A LA EDICION RUSA

Las matemáticas, surgidas en la Antigüedad por necesidades de la vida cotidiana, se han convertido en un inmenso sistema de variadas y extensas disciplinas. Como las demás ciencias, reflejan las leyes del mundo material que nos rodea y sirven de potente instrumento para el conocimiento y dominio de la naturaleza. Pero el alto nivel de abstracción que caracteriza a las matemáticas trae consigo el que todas sus ramas sean relativamente inaccesibles a los no especialistas. Esta cualidad abstracta de las matemáticas dio lugar, ya en la Antigüedad, a nociones idealistas sobre su independencia respecto del mundo material.

En la preparación del presente volumen, la finalidad perseguida por los autores ha sido la de familiarizar a un círculo suficientemente amplio de la intelligentsia soviética con las diversas disciplinas matemáticas, su contenido y métodos, los fundamentos sobre los que descansan y los caminos a lo largo de los cuales se han desarrollado.

El bagaje mínimo de conocimientos matemáticos que precisa el lector de este libro son las matemáticas a nivel de enseñanza media, aunque los volúmenes difieren en lo que toca a la accesibilidad del material tratado. El lector que desee iniciarse en los elementos de las

matemáticas superiores puede leer con provecho algunos de los primeros capítulos; sin embargo, para una completa comprensión de las partes subsiguientes le será necesario recurrir a los correspondientes libros de texto. El libro en su totalidad podrá ser asimilado de manera fundamental solamente por aquellos lectores que ya tengan algún conocimiento de las aplicaciones del análisis matemático, es decir, del cálculo diferencial y del cálculo integral. Para tales lectores —profesores de matemáticas, ingeniería y ciencias naturales— será particularmente importante la lectura de aquellos capítulos en los que se introducen las ramas más recientes de las matemáticas.

Naturalmente, dentro de los límites de un solo libro no ha sido posible agotar toda la riqueza ni siquiera de los resultados más fundamentales de la investigación matemática; una cierta libertad en la elección del material ha sido inevitable. Pese a ello, en líneas generales, el presente libro dará una idea de la situación actual de las matemáticas, sus orígenes y su probable desarrollo futuro. Por esta razón, la obra está pensada también hasta cierto punto para personas ya familiarizadas con la mayoría de los temas contenidos en ella, lo cual quizá contribuya a eliminar esa estrechez de miras que se descubre ocasionalmente en algunos de nuestros jóvenes matemáticos.

Los distintos capítulos del libro están escritos por diversos autores, cuyos nombres se dan en el índice; no obstante, el libro en su conjunto ha sido escrito en colaboración. El plan general, la elección del material, las sucesivas versiones de capítulos concretos, todo ello fue sometido a discusión general, introduciendo mejoras aquí y allá tras un animado cambio de opiniones. Matemáticos de distintas ciudades de la Unión Soviética tuvieron la oportunidad, mediante discusiones organizadas, de hacer numerosas y valiosas observaciones concernientes a la versión original del texto. Sus opiniones y sugerencias han sido tenidas en cuenta por los autores.

Los autores de algunos capítulos también tomaron parte directa en la redacción de la versión final de otros capítulos: la introducción del capítulo 2 fue escrita esencialmente por B. N. Delone, mientras que D. K. Fadiev desempeñó un activo papel en la preparación de los capítulos 4 y 18. Parte del trabajo fue también realizado por personas distintas de los autores de los diferentes capítulos: la sección 6 del capítulo 6 fue redactada por O. A. Ladyzenskaia, la sección 5 del capítulo 10 por A. G. Postnikov; sobre el texto del capítulo 5 trabajó O. A. Oleinik, y sobre el del capítulo 11, Ju. V. Prohorov.

Ciertas secciones de los capítulos 1, 2, 7, 15 fueron escritas por V. A. Zalgaller. La edición del texto final fue realizada por V. A. Zalgaller y V. S. Videnski con la colaboración de T. V. Rogozkinaia y A. P. Leonovaia.

La mayor parte de las ilustraciones fueron preparadas por E. P. Senkin.

Moscú
1956

COMITÉ DE EDITORES

Parte 5

Capítulo 14

TEORIA DE FUNCIONES DE UNA VARIABLE REAL

§1. *Introducción*

A finales del siglo XVIII y comienzos del XIX el cálculo diferencial e integral estaba esencialmente terminado. Desde ese momento (a lo largo de todo el siglo XIX), los matemáticos se dedicaron a la construcción de sus distintas ramas, al descubrimiento de nuevos resultados y al desarrollo de nuevos campos de aplicación a la mecánica, la astronomía y la tecnología. Ahora ya se podía repasar los resultados obtenidos, estudiarlos sistemáticamente y profundizar en los conceptos básicos del análisis. Y entonces se vio que los fundamentos de esta disciplina no eran tan sólidos como parecía.

Ya en el siglo XVIII los más grandes matemáticos no se ponían de acuerdo acerca de lo que era una función. El desacuerdo se tradujo en prolongadas controversias sobre si era ésta o aquella la solución de un problema, o si era o no correcto un resultado matemático concreto. Poco a poco se fue viendo que no era éste el único concepto que necesitaba de una mayor precisión. Una comprensión inadecuada del significado de la continuidad y de las propiedades de las funciones continuas condujo a resultados erróneos: por ejemplo, que una función continua es siempre diferenciable. La matemática comenzó a operar con funciones tan complicadas que resul-

taba ya imposible confiar en la intuición. Nació así la necesidad real de poner orden en los conceptos fundamentales del análisis.

El primer intento serio en esta dirección fue realizado por Lagrange, seguido después por Cauchy. Cauchy afinó las definiciones de límite, continuidad e integral, que se conservan hasta el presente, y las hizo de uso común. Aproximadamente al mismo tiempo, el matemático checo Bolzano realizó un riguroso estudio de las propiedades básicas de las funciones continuas.

Consideremos con más detalle estas propiedades. Supongamos dada una función $f(x)$ en un intervalo $[a, b]$, es decir, definida para todos los números que satisfacen las desigualdades $a \leq x \leq b$. Antes se consideraba obvio que si una función tomaba valores de signos opuestos en los puntos extremos del intervalo, entonces debía tomar el valor cero en algún punto intermedio. Ahora este hecho posee una demostración rigurosa. De la misma forma se probó con todo rigor que una función continua dada sobre un intervalo toma en ciertos puntos sus valores máximo y mínimo.

El estudio de estas propiedades de las funciones continuas hizo necesaria una mayor profundización en la naturaleza de los números reales. Como resultado de esto apareció la teoría de los números reales; las propiedades básicas de la recta numérica fueron claramente formuladas.

Más adelante el desarrollo del análisis matemático exigió el estudio de funciones cada vez más «extrañas», en particular discontinuas. Las funciones discontinuas aparecieron, por ejemplo, como límites de funciones continuas, donde no se conoce *a priori* si la función límite es continua o no, así como al esquematizar procesos que presentan variaciones bruscas. De este modo surgió una nueva tarea, a saber, la generalización del aparato del análisis a las funciones discontinuas.

Riemann investigó el problema de las clases de funciones discontinuas a las que se puede extender el concepto de integral. Como resultado de este trabajo sobre los fundamentos del análisis nació una nueva disciplina matemática: la teoría de funciones de una variable real.

Mientras que el análisis matemático clásico opera esencialmente con funciones «buenas» (por ejemplo, continuas o diferenciables), la teoría de funciones de una variable real investiga clases de funciones mucho más amplias. Mientras que en el análisis matemático clásico la definición de una operación (por ejemplo, la integración) se da para funciones continuas, es característico de la teoría de fun-

ciones de una variable real descubrir para qué clase de funciones es aplicable esta definición, y cómo debe ser modificada ésta para que sea utilizable en una clase más amplia. En particular, solamente la teoría de funciones puede dar respuesta satisfactoria al problema de la longitud de una curva y de la clase de curvas para las que tiene sentido hablar de longitud.

Los fundamentos sobre los que se construye la teoría de funciones de una variable real es la *teoría de conjuntos*.

De acuerdo con esto, comenzamos la exposición con una descripción de los elementos de la teoría de conjuntos; a continuación estudiaremos los conjuntos de puntos, cerrando el capítulo con una exposición de uno de los conceptos más fundamentales de la teoría de funciones de una variable real: la integral de Lebesgue.

§2. *Conjuntos*

Todo el mundo trata a diario con diversas colecciones de objetos. Como se explicó ya en el capítulo 1, esto supuso el desarrollo del concepto *número* y más tarde el de *conjunto*, que es uno de los conceptos básicos y primitivos de las matemáticas y no se presta a una definición precisa. Las siguientes observaciones sólo quieren ilustrar lo que es un conjunto, pero no pretenden servir de definición.

Conjunto es el nombre que designa un agregado, colección o totalidad de cosas que se reúnen bajo un cierto criterio o de acuerdo con una cierta regla. El concepto de conjunto surge por abstracción. Al considerar una colección de objetos como un conjunto, no tenemos en cuenta las conexiones y relaciones entre los objetos que lo forman, pero conservamos sus características individuales. Así, el conjunto formado por cinco monedas y el formado por cinco manzanas son diferentes. Pero el conjunto formado por cinco monedas colocadas en círculo y el conjunto formado por las mismas cinco monedas en fila es uno y el mismo conjunto.

Demos algunos ejemplos de conjuntos. Podemos hablar de los granos que forman un montón de arena, del conjunto de los planetas que forman nuestro sistema solar, del conjunto de todas las personas que están en una casa en un momento dado o del conjunto de todas las páginas de este libro. En matemáticas encontramos constantemente conjuntos: el conjunto de las raíces de una ecuación dada, el conjunto de todos los números naturales, el conjunto de los puntos de una recta, etc.

La disciplina matemática que estudia las propiedades generales de los conjuntos es la *teoría de conjuntos*. Esta disciplina se comenzó a desarrollar rigurosamente a finales del siglo XIX y comienzos del XX. El fundador de la teoría científica de conjuntos es el matemático alemán G. Cantor.

El trabajo de Cantor sobre la teoría de conjuntos nació al estudiar cuestiones de convergencia de series trigonométricas. Fenómeno muy común: muy a menudo el trabajo sobre problemas matemáticos concretos conduce a la construcción de una teoría sumamente abstracta y general. El valor de tales construcciones abstractas radica en el hecho de que a la postre resultan estar relacionadas, no sólo con el problema concreto del que surgen, sino con otra serie de problemas. En particular, este es el caso de la teoría de conjuntos. Las ideas y conceptos de la teoría de conjuntos irrumpieron literalmente en todas las ramas de las matemáticas y cambiaron su faz por completo. Por esta razón es imposible dar hoy una imagen adecuada de las matemáticas sin comenzar por la teoría de conjuntos. Para la teoría de funciones de una variable real tiene una importancia particularmente grande.

Un conjunto puede considerarse bien definido cuando se puede decir de cada objeto si pertenece o no al conjunto. De otra forma, un conjunto está completamente determinado por los objetos que pertenecen a él. Si un conjunto M está formado por los objetos $a, b, c, \dots$, y no otros, escribimos

$$M = \{a, b, c, \dots\}$$

Los objetos que forman un cierto conjunto se dice que son sus *elementos*. El hecho de que un objeto m sea un elemento de M se escribe así:

$$m \in M$$

y se lee: « m pertenece a M » o « m es un elemento de M ». Si un objeto n no pertenece al conjunto M , entonces se escribe: $n \notin M$ *. Cada objeto puede solamente ser un elemento de un conjunto dado; es decir, todos los elementos de un conjunto son distintos unos de otros.

Los elementos de un conjunto pueden a su vez ser conjuntos; sin embargo, para evitar contradicciones, es conveniente postular que un conjunto M no puede ser uno de sus propios elementos, $M \notin M$.

* También es corriente escribir $\notin$ en lugar de ϵ . (*N. del T.*)

El conjunto que no contiene ningún elemento se llama *conjunto vacío*. Por ejemplo, el conjunto de las raíces reales de la ecuación

$$x^2 + 1 = 0$$

es vacío. Desde ahora el conjunto vacío se denotará por ϕ .

Si para dos conjuntos M y N cada elemento x de M lo es también de N , decimos que M es parte de N , que M es un subconjunto de N , o que M está contenido en N ; esto se escribe de la forma

$$M \subseteq N \quad \text{o} \quad N \supseteq M.$$

Por ejemplo, el conjunto $M = \{1, 2\}$ es parte del conjunto $N = \{1, 2, 3\}$.

Evidentemente tenemos $M \subseteq N$. Es conveniente considerar el conjunto vacío como parte de cualquier conjunto.

Dos conjuntos son *iguales* si constan de los mismos elementos. Por ejemplo, el conjunto de las raíces de la ecuación $x^2 - 3x + 2 = 0$ y el conjunto $M = \{1, 2\}$ son iguales.

Definimos ahora las reglas de las *operaciones* con conjuntos.

Unión o suma

Supongamos que $M, N, P, \dots$ son conjuntos. La unión o suma de ellos es el conjunto X formado por todos los elementos que pertenecen al menos a uno de los «sumandos» $M, N, P, \dots$

$$X = M + N + P + \dots$$

Ahora bien, si un elemento x pertenece a varios sumandos, pertenece a la suma X una sola vez. Está claro entonces que

$$M + M = M$$

y que si $M \subseteq N$, entonces

$$M + N = N.$$

Intersección

La intersección o parte común de los conjuntos $M, N, P, \dots$ es el conjunto Y que consta de los elementos que pertenecen simultáneamente a todos los conjuntos $M, N, P, \dots$

Evidentemente, $M \cdot M = M$; y si $M \subseteq N$, entonces $M \cdot N = M$.

Si la intersección de dos conjuntos M y N es vacía, $M \cdot N = \phi$, decimos que los conjuntos son *disjuntos*.

Como notación para la suma e intersección de conjuntos, usamos también los símbolos $\sum$ y $\prod$. Así,

$$E = \sum E_i$$

es la suma de los conjuntos E_i , y

$$F = \prod E_i,$$

su intersección.

Recomendamos al lector que demuestre que la suma y la intersección de conjuntos están relacionadas por la ley distributiva usual

$$M(N + P) = MN + MP,$$

y también por la ley

$$M + NP = (M + N)(M + P).$$

Diferencia

La diferencia de dos conjuntos M y N es el conjunto Z de todos los elementos de M que no pertenecen a N ,

$$Z = M - N.$$

Si $N \subseteq M$, entonces la diferencia $Z = M - N$ se llama también *complemento* de N en M .

No es difícil demostrar que siempre

$$M(N - P) = MN - MP$$

y

$$(M - N) + MN = M.$$

Así, las reglas de las operaciones con conjuntos difieren de las reglas usuales de la aritmética.

Conjuntos finitos e infinitos

Los conjuntos que constan de un número finito de elementos se llaman conjuntos finitos. Si el número de elementos no está acotado, el conjunto se llama infinito. Por ejemplo, el conjunto de todos los números naturales es infinito.

Consideremos dos conjuntos arbitrarios M y N , y comparemos el número de sus elementos.

Si el conjunto M es finito, entonces la colección de sus elementos está caracterizada por un cierto número natural, a saber, el número de sus elementos. Así, pues, al objeto de comparar el número de elementos de M y N , es suficiente contar el número de elementos de M y de N y comparar los números así obtenidos. También es natural considerar que si uno de los conjuntos es finito y el otro es infinito, entonces el conjunto infinito posee más elementos que el finito.

Sin embargo, si ambos conjuntos son infinitos, un simple cómputo de sus elementos no conduce a nada. Surge así el siguiente problema: ¿tienen todos los conjuntos infinitos el mismo número de elementos o, por el contrario, existen conjuntos infinitos con un número mayor o menor de elementos? Si lo último es cierto, ¿cómo se pueden comparar los conjuntos infinitos? Centremos ahora nuestra atención sobre estos problemas.

Correspondencias uno a uno

Sean de nuevo M y N dos conjuntos finitos. ¿Cómo saber qué conjunto tiene más elementos sin contar el número de ellos? A este fin formemos *parejas*, con un elemento de M y otro de N . Entonces, si para algún elemento de M no existe otro de N con el que formar pareja, M tiene más elementos que N . Ilustremos este razonamiento con un ejemplo.

Supongamos que en una habitación hay un cierto número de personas y un cierto número de sillas. Para saber si hay más personas que sillas, o al revés, basta rogar a las personas que se sienten. Si alguna persona queda sin asiento, esto significa que hay más personas; y si todas las personas tienen sitio y todas las plazas están ocupadas, entonces hay tantas personas como sillas. Este método de comparar el número de elementos de dos conjuntos tiene la ventaja, sobre el recuento directo, de que es aplicable sin modificaciones esenciales a conjuntos infinitos.

Consideremos el conjunto de todos los números naturales

$$M = \{1, 2, 3, 4, \dots\}$$

y el conjunto de todos los números pares

$$N = \{2, 4, 6, 8, \dots\}.$$

¿Qué conjunto contiene más elementos? A primera vista parece que el primero. Sin embargo, podemos formar parejas de elementos como indica la tabla 1:

TABLA 1

M	1	2	3	4	...
N	2	4	6	8	...

Ningún elemento de M ni de N queda sin pareja. Ciertamente podíamos haber formado parejas como en la tabla 2:

TABLA 2

M	1	2	3	4	5	6	7	...
N	—	2	—	4	—	6	—	...

Entonces muchos elementos de M quedan sin pareja. Por otra parte, podíamos haber formado pares como en la tabla 3:

TABLA 3

M	—	1	—	2	—	3	—	...
N	2	4	6	8	10	12	14	...

Ahora muchos elementos de N quedan sin pareja.

Por tanto, si los conjuntos A y B son infinitos, entonces métodos distintos de formación de parejas conducen a resultados también distintos. Si existe un método de formación de parejas de forma que todos los elementos de A y B queden emparejados, decimos que entre A y B se puede establecer una correspondencia uno a uno. Por ejemplo, podemos establecer una correspondencia uno a uno, como se deduce de la tabla 1, entre los conjuntos M y N .

Si entre dos conjuntos A y B podemos establecer una correspondencia uno a uno, se dice que tienen el *mismo* número de elementos. Si *cualquiera que sea* el método de formación de parejas siempre queda algún elemento de A desemparejado, decimos que A contiene más elementos que B , o que A tiene mayor *cardinalidad* que B .

Así, hemos obtenido respuesta a una de las cuestiones planteadas anteriormente: cómo comparar el número de elementos en conjun-

tos infinitos. Si embargo, queda sin contestar la otra pregunta: ¿existen conjuntos infinitos con distinta cardinalidad? Con el fin de dar respuesta a esta cuestión estudiemos algunos tipos simples de conjuntos infinitos.

Conjuntos numerables

Si podemos encontrar una correspondencia uno a uno entre un conjunto dado A y el conjunto de los números naturales

$$Z = \{1, 2, 3, \dots\},$$

decimos que A es *numerable*. De otra forma, un conjunto es numerable si todos sus elementos pueden ser enumerados por medio de los números naturales, es decir, escritos como sigue en forma de sucesión

$$a_1, a_2, a_3, \dots, a_n, \dots$$

La tabla 1 muestra que el conjunto de los números pares es numerable (los números de arriba pueden considerarse subíndices de los correspondientes de abajo).

Los conjuntos numerables son, por decirlo así, los conjuntos infinitos más pequeños: todo conjunto infinito contiene un subconjunto numerable.

Si dos conjuntos finitos no vacíos son disjuntos, su suma contiene más elementos que cada uno de los sumandos. Para conjuntos infinitos esto no es cierto siempre. Por ejemplo, sea P el conjunto de todos los números pares; I , el conjunto de todos los impares, y Z , el conjunto de los números naturales. Como muestra la tabla 4, los conjuntos P e I son numerables. Sin embargo, el conjunto $Z = P + I$ es también numerable.

TABLA 4

P	2	4	6	8	...
I	1	3	5	7	...
Z	1	2	3	4	...

La violación de la regla «el todo es mayor que las partes» en conjuntos infinitos muestra que las propiedades de los conjuntos infinitos difieren *cualitativamente* de las de los conjuntos finitos. La

transición de finito a infinito procede de completo acuerdo con el bien conocido principio dialéctico de la variación cualitativa de propiedades.

Probemos ahora que *el conjunto de los números racionales es numerable*. Para ello escribimos los números racionales en la tabla 5

TABLA 5

(1)	(2)	(3)	(4)	(5)	(6)	
1	2	3	4	5	6	
0	-1	-2	-3	-4	-5	
$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$	$\frac{9}{2}$	$\frac{11}{2}$	
$-\frac{1}{2}$	$-\frac{3}{2}$	$-\frac{5}{2}$	$-\frac{7}{2}$	$-\frac{9}{2}$	$-\frac{11}{2}$	
$\frac{1}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{5}{3}$	$\frac{7}{3}$	$\frac{8}{3}$	...
$-\frac{1}{3}$	$-\frac{2}{3}$	$-\frac{4}{3}$	$-\frac{5}{3}$	$-\frac{7}{3}$	$-\frac{8}{3}$	...
...	...	...	...	...	...	...

Aquí están colocados en la primera fila, en orden creciente, los números naturales; en la segunda, el cero y todos los números negativos en orden decreciente; en la tercera, las fracciones reducidas positivas de denominador dos en orden creciente; en la cuarta, las fracciones negativas de denominador dos en orden decreciente, etcétera. Es claro que cada número racional aparece una sola vez en esta tabla. Enumeremos ahora todos los números de la tabla en el orden indicado por las flechas. Así, todos los números racionales quedan ordenados en una única sucesión:

Número de lugar ocupado por el número racional	1	2	3	4	5	6	7	8	9	...
Números racionales	1,	2,	0,	3,	-1,	$\frac{1}{2}$,	4,	-2,	$\frac{3}{2}$,	...

De esta forma hemos establecido una correspondencia uno a uno entre los números racionales y los números naturales. Por tanto, el conjunto de los números racionales es numerable.

Conjuntos con el número cardinal del continuo

Si podemos establecer una correspondencia uno a uno entre los elementos de un conjunto M y los puntos del intervalo $0 \leq x \leq 1$, decimos que el conjunto M tiene el *número cardinal del continuo*. En particular, por definición, el conjunto de puntos del segmento $0 \leq x \leq 1$ tiene el número cardinal del continuo.

De la figura 1 se deduce que el conjunto de puntos de cualquier intervalo AB tiene el número cardinal del continuo. La correspondencia uno a uno ha sido establecida geométricamente por medio de una proyección.

No es complicado demostrar que el conjunto de puntos de cualquier intervalo abierto $a < x < b$ y de toda la recta numérica tienen el número cardinal del continuo.

De interés considerablemente mayor es el siguiente hecho: el conjunto de los puntos del cuadrado $0 \leq x \leq 1$, $0 \leq y \leq 1$, tiene el número cardinal del continuo. Hablando imprecisamente: tantos puntos hay en el cuadrado como en el segmento.

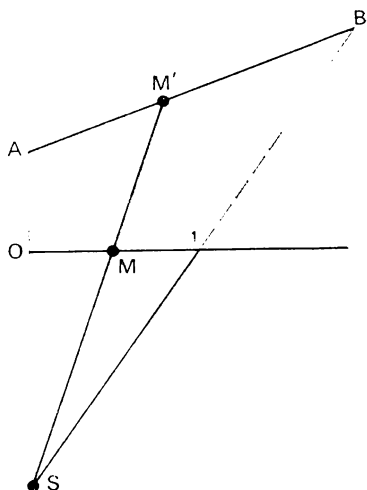


FIG. 1

§3. Números reales¹

El desarrollo del concepto de número ha sido descrito con detalle en el capítulo 1. Daremos ahora al lector una breve reseña de las teorías que sobre los números reales surgieron en el siglo XIX en relación con la fundamentación de los conceptos básicos del análisis.

¹ En la redacción de esta sección contamos con el valioso consejo de A. N. Kolmogorov.

Números racionales

Suponemos familiarizado al lector con las propiedades de los números racionales. Sin entrar en detalles, las recordaremos brevemente. Los números racionales, es decir, los números de la forma m/n , donde m y n son enteros y $n \neq 0$, forman un conjunto de números en el que se definen dos operaciones, la adición y la multiplicación. Estas operaciones están sujetas a un cierto número de leyes (axiomas). En lo que sigue, $a, b, c, \dots$ denotan números racionales.

I. Axiomas de la adición

1. $a + b = b + a$ (conmutatividad)
2. $a + (b + c) = (a + b) + c$ (asociatividad)
3. La ecuación

$$a + x = b$$

tiene una única solución (existencia de la operación inversa).

De estos axiomas se sigue inmediatamente que la expresión $a + b + c$ tiene un significado único; que existe un número racional 0 (cero) para el cual $a + 0 = a$; y que la adición tiene una operación inversa (sustracción) de forma que la expresión $b - a$ tiene significado.

Así, pues, desde el punto de vista algebraico, los números racionales forman un grupo conmutativo para la operación de adición.

II. Axiomas de la multiplicación

1. $ab = ba$ (conmutatividad)
2. $a(bc) = (ab)c$ (asociatividad)
3. La ecuación

$$ay = b,$$

siendo $a \neq 0$, tiene una única solución (existencia de operación inversa).

De estos axiomas se sigue que la expresión abc tiene sentido; que existe un número racional 1 para el cual $a \cdot 1 = a$; y que para

números racionales distintos del 0 la operación inversa (división) existe. Los números racionales, exceptuado el 0, forman un grupo conmutativo para la multiplicación.

III. Axioma de distributividad

$$1. (a + b)c = ac + bc$$

Los axiomas I-III indican que para las operaciones de adición y multiplicación los números racionales forman un *cuerpo algebraico*.

IV. Axiomas de orden

1. Para cualesquiera dos números racionales es verdadera una y sólo una de las tres siguientes relaciones: o bien $a < b$, o bien $a > b$, o bien $a = b$.
2. Si $a < b$, y $b < c$, entonces $a < c$.
3. Si $a < b$, entonces $a + c < b + c$ (monotonía de la adición).
4. Si $a < b$ y $c > 0$, entonces $ac < bc$ (monotonía de la multiplicación por $c > 0$).

Todos estos axiomas hacen del conjunto de los números racionales un *cuerpo ordenado*.

Además de los números racionales existen otros sistemas de objetos que verifican estos axiomas y que son, por tanto, cuerpos ordenados.

Mencionamos dos importantes propiedades de los números racionales.

Densidad. Cualesquiera que sean a y b , $a < b$, existe un c tal que $a < c < b$.

Numerabilidad. El conjunto de los números racionales es numerable (ver §2).

Medida de cantidades

La insuficiencia de los números racionales en matemáticas resulta evidente al tratar problemas tan importantes como la medida de cantidades. Como uno de los ejemplos más simples vamos a considerar el problema de la medida de intervalos.

Tomemos una recta sobre la que se ha definido un sentido, un

origen (el punto 0) y una unidad de escala. Podemos entonces construir segmentos OA de longitud $1/2$, $1/3$, $2/3$, $-1/3$, etc. Más generalmente, podemos asociar a cada número racional a un punto A sobre la recta, a saber, el punto de coordenada $x = a$. En este caso el número a determina la longitud del segmento OA . Sin embargo, no es cierto que todo segmento tenga un número racional como medida. Los griegos, por ejemplo, sabían ya que la diagonal de un cuadrado de lado unidad no puede ser medida con un número racional. La salida natural de esta situación es establecer una correspondencia uno a uno entre números y longitudes, es decir, realizar una nueva extensión del concepto de número.

Números reales

Hemos llegado a la conclusión de que los números racionales, exclusivamente, son insuficientes para medir cantidades y que el concepto de número debe ser ampliado de forma que exista una correspondencia uno a uno entre los números y los puntos de una recta. Teniendo esto en cuenta tratemos de ver si es posible determinar la posición de un punto *arbitrario* sobre la recta por medio solamente de puntos racionales. Una construcción similar en el campo de los números racionales nos conducirá al concepto de número real.

Sea α un punto arbitrario sobre la recta. Podemos dividir los puntos racionales en dos clases: en una ponemos los puntos α que estén a la izquierda de α y en otra los que se hallen a la derecha. En cuanto al punto α (si es racional), podemos ponerlo en una u otra clase. Una división de este tipo en los puntos racionales se suele llamar *cortadura*. Dos cortaduras son iguales si sus clases son iguales (salvo quizá en un único punto). No es difícil probar que dos puntos distintos α y β dan lugar a distintas cortaduras, pues, como los puntos racionales son densos en toda la recta, podemos encontrar puntos racionales r_1 y r_2 situados estrictamente entre α y β . Así, en una de las cortaduras los puntos r_1 y r_2 estarán en la clase de la derecha, y en la otra, en la de la izquierda.

Así, cada punto de la recta determina una cortadura en el campo de los puntos racionales, y cortaduras diferentes corresponden a puntos diferentes. Es muy importante que una cortadura pueda ser definida de otras formas sin hacer referencia al número α . Para ello definimos una cortadura en el campo de los puntos racionales como una división de éstos en dos conjuntos A y B no vacíos y disjuntos tales que $a < b$ para todo $a \in A$, $b \in B$. Según esta definición pode-

mos asignar a cada cortadura de manera única el punto (frontera) que determinan. Dicho con otras palabras, mediante cortaduras en el campo de los puntos racionales podemos determinar *cualquier* punto de la recta. La construcción que hemos desarrollado fue propuesta por el matemático alemán R. Dedekind y se conoce con el nombre de *cortadura de Dedekind*.

Las cortaduras son el único método posible de determinar la posición de un punto arbitrario mediante puntos racionales. Más próximo al método usual de medir en la práctica está el siguiente método, debido a G. Cantor. Sea de nuevo α un punto cualquiera sobre la recta. Podemos tomar entonces dos puntos arbitrariamente próximos de forma que α quede entre ellos. Los puntos a y b determinan la posición de α aproximadamente. Imaginemos este proceso de determinación aproximada repetido indefinidamente, de forma que la precisión aumente cada vez más. Obtenemos así un sistema de intervalos $[a_n, b_n]$ con los extremos en puntos racionales tales que $[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n]$ y $b_n - a_n \rightarrow 0 (n \rightarrow \infty)$. Un sistema de intervalos de este tipo recibe el nombre de *sistema de intervalos encajados*. Está claro que un sistema de este tipo determina la posición de α de manera única.

Por medio de construcciones similares en el cuerpo de los números racionales se llega a definir los números reales. A continuación definiremos las operaciones entre ellos y veremos que satisfacen los mismos axiomas que las operaciones con números racionales. A cada punto de la recta corresponde un número real y viceversa. Por eso, al conjunto de los números reales se le denomina a menudo recta numérica.

Principios de continuidad

Existen diferencias esenciales entre el conjunto de los números racionales y el de los reales. De hecho, el conjunto de los números reales tiene una serie de propiedades de *continuidad* que no tiene el conjunto de los números racionales. Habitualmente se denominan principios de continuidad. Enumeraremos los más importantes.

Principio de Dedekind. Si el conjunto de los números reales se divide en dos conjuntos no vacíos X e Y sin elementos en común, de forma que para $x \in X$, $y \in Y$ arbitrarios se cumpla $x < y$, existe entonces un único número ξ (la frontera) para el cual $x \leq \xi \leq y$, cualesquiera que sean $x \in X$, $y \in Y$.

El conjunto de los números reales x tales que $a \leq x \leq b$ se dice

que es un intervalo de la recta numérica y se denota por $[a, b]$. Un sistema de intervalos $[a_n, b_n]$ se llama *encajado* si $[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n]$ y $b_n - a_n \rightarrow 0$ ($n \rightarrow \infty$).

Principio de Cantor. Dado un sistema de intervalos encajados $[a_n, b_n]$ existe un único número real ξ que pertenece a cada uno de los intervalos.

Principio de Weierstrass. Si una sucesión no decreciente de números reales está acotada superiormente, entonces es convergente.

Decimos que una sucesión de números reales $\{x_n\}$ es una sucesión *fundamental* si para todo $\epsilon > 0$ podemos encontrar un número natural N tal que para todo $n > N$ y todo número natural p

$$|x_{n+p} - x_n| < \epsilon.$$

Principio de Cauchy. Toda sucesión fundamental de números reales converge.

Puesto que no hemos dado una construcción precisa de los números reales, no estamos en condiciones de probar estos principios. Nuestro objetivo ahora consiste en investigar cómo están relacionados. Supongamos que uno de los principios de continuidad es cierto y examinemos si los demás se siguen de él.

El resultado global a que llegaremos es que los principios de continuidad son equivalentes.

Decimos que b es el extremo superior (mínimo de las cotas superiores) del conjunto E , y escribimos

$$b = \sup E,$$

si 1) $x \leq b$ para todo $x \in E$, y 2) no existe un número $b' < b$ con la misma propiedad.

Mostremos que la siguiente proposición se sigue del principio de Dedekind: todo conjunto E (no vacío) de números reales acotado superiormente tiene extremo superior. Dividimos los números reales en dos clases X e Y según el siguiente criterio: ponemos $x \in X$ si existe un $a \in A$ tal que $a \geq x$, y ponemos $y \in Y$ si tenemos $a < y$ para todo $a \in A$. Es fácil verificar que esto es una cortadura en el cuerpo de los números reales. Por el principio de Dedekind tiene una frontera ξ ; esta frontera es el extremo superior de E .

Probemos ahora que el principio de Weierstrass se sigue del principio de Dedekind. Sea $\{x_n\}$ una sucesión no decreciente de

números reales acotada superiormente. Por lo que hemos visto, tiene extremos superior ξ . Por definición de cota superior $x_n \leq \xi$ ($n = 1, 2, \dots$); para cada $\epsilon > 0$ podemos encontrar un índice n_0 tal que $x_{n_0} > \xi - \epsilon$. Puesto que la sucesión $\{x_n\}$ es monótona, esto implica que $\xi - \epsilon < x_n \leq \xi$ para todo $n > n_0$, es decir, la sucesión $\{x_n\}$ converge al límite ξ .

Para probar la relación inversa entre los principios de Dedekind y Weierstrass, demostremos que este último implica el siguiente principio:

Principio de Arquímedes. Cualesquiera que sean los números reales $a > 0$ y b , podemos encontrar un número natural n tal que $na > b$.

Este principio significa que para todo número real b la sucesión $\{b/n\}$ ($n = 1, 2, \dots$) converge a 0.

Supongamos que se cumple el principio de Weierstrass pero no el de Arquímedes. Esto último significa que existe un número $a > 0$ tal que la sucesión $x_n = na$ está acotada. Además es creciente. Por el principio de Weierstrass tiene un límite ξ . Por tanto, el intervalo $[\xi - a/2, \xi]$ contiene algún punto $x_n = na$ de nuestra sucesión. Pero entonces $x_{n+1} = (n+1)a > \xi$, y esto contradice el hecho de que ξ es el extremo superior de $\{x_n\}$.

El principio de Weierstrass implica el principio de Dedekind. Supongamos los números reales divididos en dos conjuntos disjuntos X e Y de modo que $x < y$ para todo $x \in X$, $y \in Y$. Probaremos que esta cortadura tiene una única frontera ξ . Sea m un entero y n un número natural. Denotamos por x_n el mayor elemento de la forma $m/2^n \in X$ tal que $x_n + 1/2^n \in Y$. Puesto que el conjunto de elementos de la forma $m/2^n$ está contenido en el conjunto de elementos de la forma $m/2^{n+1}$, tenemos $x_n \leq x_{n+1}$. Además, la sucesión $\{x_n\}$ está acotada (por ejemplo, por el número $x_1 + 1/2$). Por tanto, según el principio de Weierstrass tiene un límite ξ . Mostremos que ξ es la frontera de nuestra cortadura. Si $x < \xi$, entonces $x \in X$. Y si $y > \xi$, entonces $y \in Y$, porque según el principio de Arquímedes podemos encontrar un número n tal que $1/2^n < y - \xi = a$. Pero $x_n < \xi$, $x_n + 1/2^n \in Y$, y entonces $y = \xi + a > x_n + 1/2^n$; por consiguiente $y \in Y$.

Se puede probar también que los principios de Cantor y Cauchy son equivalentes. Sin embargo, la validez del principio de Cauchy no implica la del de Dedekind. Esta afirmación debe ser tomada en el siguiente sentido: existen cuerpos ordenados en los que se verifica el principio de Cauchy pero no el de Dedekind. Si suponemos de

antemano que se cumple el principio de Arquímedes, entonces los cuatro principios son equivalentes.

La no numerabilidad del continuo

Probemos que el conjunto de los puntos del intervalo $0 \leq x \leq 1$ no es numerable. Daremos una demostración indirecta. Supongamos que el conjunto de los puntos del intervalo $0 \leq x \leq 1$ es numerable. Entonces podemos escribir todos los puntos del segmento en forma de sucesión

$$x_1, x_2, \dots, x_n, \dots \quad (1)$$

Tomamos en $[0, 1]$ un intervalo σ_1 tal que su longitud sea menor que 1 y no contenga el punto x_1 . Un intervalo así se puede encontrar fácilmente. Ahora, dentro de σ_1 tomamos un intervalo σ_2 tal que su longitud sea menor que $1/2$ y no contenga los puntos x_1, x_2 . En general, una vez tomado el intervalo σ_{n-1} se escoge el intervalo σ_n de forma que su longitud sea menor que $1/n$, no contenga los puntos $x_1, x_2, \dots, x_n$, y esté contenido en σ_{n-1} . De esta forma hemos construido una sucesión infinita de intervalos

$$\sigma_1, \sigma_2, \dots, \sigma_n, \dots,$$

tales que cada uno está contenido en el precedente y su longitud tiende a cero cuando n crece. Por el principio de Cantor, en el intervalo $[0, 1]$ hay un único punto x que pertenece a todos los intervalos σ_n . Puesto que, por hipótesis, todos los puntos de $[0, 1]$ están en la sucesión (1), el punto x que es común a todos los σ_n coincide con algún punto x_m de la sucesión. Pero por construcción σ_m no contiene a x_m , de modo que $x \neq x_m$. Así, pues, hemos llegado a una contradicción. Por consiguiente, la hipótesis inicial, según la cual el conjunto de los puntos de $[0, 1]$ es numerable, es falsa; el intervalo $[0, 1]$ no es numerable. Esto es lo que tratábamos de probar.

Este teorema demuestra que existen cardinalidades infinitas distintas, dando así una respuesta positiva a la primera cuestión planteada.

§4. *Conjuntos de puntos*

En la sección precedente hemos encontrado ya conjuntos cuyos elementos eran puntos. En particular, hemos considerado el conjunto de los puntos de un intervalo cualquiera y el conjunto de los

puntos (x, y) del cuadrado $0 \leq x \leq 1, 0 \leq y \leq 1$. Vamos a estudiar ahora con más detalle las propiedades de tales conjuntos.

Un conjunto cuyos elementos son puntos se denomina un *conjunto de puntos*. Así, podemos hablar de conjuntos de puntos de la recta, del plano o de un espacio arbitrario. Para mayor sencillez nos limitaremos al estudio de conjuntos de puntos *de la recta*.

Existe una estrecha conexión entre los números reales y los puntos de la recta: a cada número real se le puede asociar un punto de la recta, y viceversa. Por tanto, al hablar de conjuntos de puntos podemos también incluir los conjuntos de números reales. Es más, para definir un conjunto de puntos de la recta daremos, por regla general, el conjunto de sus coordenadas.

Los conjuntos de puntos (y, en particular los conjuntos de puntos de la recta) tienen propiedades especiales que los distinguen de los conjuntos arbitrarios, dando lugar a una disciplina matemática con contenido propio. En primer lugar, tiene sentido hablar de *distancia* entre dos puntos. Además podemos establecer una relación de *orden* (izquierda, derecha) entre los puntos de la recta; de acuerdo con esto se dice que un conjunto de puntos de la recta es un conjunto *ordenado*. Finalmente, como ya mencionamos anteriormente, el principio de Cantor se cumple para la recta; esta propiedad suele denominarse *completitud* de la recta.

Introducimos ahora una notación nueva para los conjuntos más simples.

Un *intervalo* $[a, b]$ es un conjunto de puntos cuyas coordenadas satisfacen la desigualdad $a \leq x \leq b$.

Un *intervalo abierto* (a, b) es un conjunto de puntos cuyas coordenadas satisfacen las condiciones $a < x < b$.

Los *semiintervalos* $[a, b)$ y $(a, b]$ están definidos por las condiciones $a \leq x < b$ y $a < x \leq b$, respectivamente.

Los intervalos abiertos y los semiintervalos pueden ser *impropios*. Así, $(-\infty, \infty)$ denota toda la recta real, y $(-\infty, b]$ el conjunto de los puntos que satisfacen $x \leq b$.

Comencemos con la enumeración de las distintas posibilidades para la situación de un conjunto (considerado como un todo) sobre la recta.

Conjuntos acotados y no acotados

Un conjunto E de puntos sobre la recta puede constar, o bien de puntos cuya distancia al origen de coordenadas no excede de un cierto número positivo, o bien de puntos arbitrariamente alejados

del origen. En el primer caso se dice que el conjunto E es acotado, y en el segundo, que es no acotado. Un ejemplo de conjunto acotado es el formado por los puntos del intervalo $[0, 1]$, y un ejemplo de conjunto no acotado, el formado por los puntos de coordenadas enteras.

Es fácil ver que, si a es un punto fijo de la recta, un conjunto E es acotado si, y sólo si, las distancias de a a un punto arbitrario $x \in E$ no exceden de un cierto número positivo.

Conjuntos acotados inferior y superiormente

Sea E un conjunto de puntos de la recta. Si existe un punto A sobre la recta tal que todo punto $x \in E$ está a la izquierda de A , decimos que E está acotado *superiormente*. Análogamente, si existe un punto a sobre la recta tal que todo $x \in E$ está a la derecha de a , decimos que E está acotado *inferiormente*. Así, el conjunto de puntos con coordenadas positivas está acotado inferiormente, y el formado por los puntos de coordenadas negativas, acotado superiormente.

Es evidente que la definición de conjunto acotado es equivalente a la siguiente: un conjunto E de puntos sobre la recta se llama acotado si está acotado superior e inferiormente. A pesar de que estas dos definiciones son similares, existe una diferencia esencial entre ellas: la primera se basa en la definición de una distancia entre los puntos de la recta, y la segunda, en el hecho de que los puntos forman un conjunto ordenado.

Podemos decir también que un conjunto es acotado si está contenido en algún intervalo $[a, b]$.

Extremo superior e inferior de un conjunto

Supongamos que un conjunto E está acotado superiormente. Existen entonces puntos A sobre la recta tales que ningún punto de E queda a su derecha. Entre todos estos puntos existe uno a la izquierda de los demás. Este punto se llama *extremo superior* de E . El *extremo inferior* se define análogamente.

Si en E existe un punto a la derecha de todos los demás puntos de E , éste es evidentemente el extremo superior. Puede suceder, sin embargo, que E no posea dicho punto.

Por ejemplo, el conjunto de puntos de coordenadas

$$\frac{0}{1}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \dots$$

está acotado superiormente, pero ninguno está a la derecha de todos los demás. En este caso el extremo a no pertenece a E , pero existen puntos de E arbitrariamente cerca de a . En el ejemplo anterior $a = 1$.

Posiciones de un punto respecto a un conjunto

Sea E un conjunto y x un punto arbitrario sobre la recta. Son posibles los siguientes casos:

1. Ni el punto x ni los puntos suficientemente próximos a x pertenecen a E .
2. El punto x no pertenece a E , pero existen puntos de E arbitrariamente próximos a x .
3. El punto x pertenece a E , pero ningún punto suficientemente próximo a x pertenece a E .
4. El punto x pertenece a E y existen puntos de E arbitrariamente próximos a x .

En el caso 1, se dice que x es *exterior* a E ; en el caso 3, que x es un *punto aislado* de E ; y en los casos 2 y 4, *punto límite* de E .

Así, si $x \notin E$, entonces x puede ser exterior a E o punto límite de E ; y si $x \in E$, entonces x puede ser punto aislado de E o punto límite.

Un punto límite puede o no pertenecer a E , y está caracterizado por la existencia de puntos de E arbitrariamente próximos a él. Dicho con otras palabras, un punto x es punto límite de E si todo intervalo abierto δ que contenga a x contiene infinitos puntos de E . El concepto de punto límite es uno de los más importantes en la teoría de conjuntos de puntos.

Si x y todos los puntos suficientemente próximos a él pertenecen a E , entonces se dice que x es un punto *interior* de E . Un punto x que no es ni interior ni exterior de E se denomina *punto frontera* de E .

A continuación damos algunos ejemplos para ilustrar estos conceptos.

Ejemplo 1

Sea E_1 formado por los puntos de coordenadas

$$1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, \dots$$

Todos los puntos son aislados; el punto 0 es un punto límite (que no pertenece a E_1), y todos los demás puntos de la recta son exteriores a E_1 .

Ejemplo 2

Sea E_2 formado por todos los puntos racionales del intervalo $[0, 1]$. Este conjunto no tiene puntos aislados; todo punto del intervalo $[0, 1]$ es un punto límite de E , y los restantes puntos de la recta son exteriores a E_2 . Entre los puntos límite los hay que pertenecen a E_2 y otros que no pertenecen a E_2 .

Ejemplo 3

Sea E_3 formado por *todos* los puntos del intervalo $[0, 1]$. Como en el ejemplo anterior, E_3 no tiene puntos aislados y todo punto de $[0, 1]$ es un punto límite. Sin embargo, a diferencia del ejemplo precedente, todos los puntos límite de E_3 pertenecen a E_3 .

Ejemplo 4

Sea E_4 formado por todos los puntos de la recta con coordenadas enteras. Todo punto de E_4 es aislado; además E_4 no tiene puntos límite.

Señalemos también que en el ejemplo 3 todos los puntos del intervalo abierto $(0, 1)$ son interiores a E_3 , y que en el ejemplo 2 todo punto del intervalo $[0, 1]$ es un punto límite de E_2 .

En los ejemplos precedentes se pone de manifiesto que un conjunto infinito puede tener puntos aislados (E_1, E_4) o no tenerlos (E_2, E_3); que puede tener puntos interiores (E_3) o no tenerlos (E_1, E_2, E_4). En cuanto a los puntos límite, sólo E_4 carece de ellos. Como demuestra el siguiente teorema, esto está relacionado con el hecho de que E_4 no está acotado.

Teorema de Bolzano-Weierstrass

Todo conjunto infinito y acotado de puntos de la recta tiene al menos un punto límite.

Probemos el teorema. Supongamos que E es un conjunto infinito y acotado de puntos de la recta. Puesto que E está acotado, está contenido en algún intervalo $[a, b]$. Dividimos el intervalo por la mitad. Como E es infinito, al menos uno de los intervalos así obtenidos contiene infinitos puntos de E . Denotemos tal intervalo por σ_1 (si las dos mitades de $[a, b]$ contienen infinitos puntos de E , σ_1 denotará, por ejemplo, la mitad de la izquierda). A continuación dividimos σ_1 en dos mitades iguales. Puesto que la parte de E contenida en σ_1 es infinita, al menos uno de los intervalos así obtenidos

contiene infinitos puntos de E . Denotamos a éste por σ_2 . Continuando este proceso de dividir un intervalo por la mitad indefinidamente y seleccionando cada vez la parte que contiene infinitos puntos de E , obtenemos una sucesión de intervalos $\sigma_1, \sigma_2, \dots, \sigma_n, \dots$. Esta sucesión tiene las tres propiedades siguientes: cada intervalo σ_{n+1} está contenido en el precedente σ_n y contiene infinitos puntos de E ; la longitud de los intervalos tiende a cero. Las dos primeras propiedades de la sucesión se deducen inmediatamente de la construcción, y para probar la última es suficiente observar que, si la longitud de $[a, b]$ es l , la longitud de σ_n es $l/2^n$. Por el principio de Cantor existe un único punto x que pertenezca a todos los σ_n . Debemos demostrar que x es un punto límite de E . Para ello basta asegurarse de que, si δ es un intervalo abierto que contiene a x , contiene también infinitos puntos de E . Puesto que cada intervalo σ_n contiene a x y su longitud tiende a cero, para un n suficientemente grande el intervalo σ_n está enteramente contenido en δ . Pero por hipótesis σ_n contiene infinitos puntos de E . Por tanto δ debe contener infinitos puntos de E . De esta forma x es un punto límite de E , y el teorema está demostrado.

Ejercicio

Probar que, si un conjunto E está acotado inferiormente y no existe ningún punto de él que quede a la izquierda de todos los demás, su extremo inferior es un punto límite de E (y no pertenece a E).

Conjuntos abiertos y cerrados

Uno de los objetivos fundamentales de la teoría de conjuntos de puntos es estudiar las propiedades de los distintos tipos de conjuntos. Familiaricemos al lector con esta teoría mediante dos ejemplos, estudiando los conjuntos llamados abiertos y cerrados.

Un conjunto se llama *cerrado* si contiene todos sus puntos límite. Si un conjunto no tiene puntos límite también será tomado como cerrado. Aparte de sus puntos límite un conjunto cerrado puede contener también puntos aislados. Un conjunto se llama *abierto* si todos sus puntos son interiores.

Demos algunos ejemplos de conjuntos cerrados y abiertos. Los intervalos $[a, b]$ son conjuntos cerrados y los intervalos (a, b) son conjuntos abiertos. Los intervalos impropios $(-\infty, b]$ y $[a, \infty)$ son cerrados, y los $(-\infty, b)$ y (a, ∞) son abiertos. La recta real es un conjunto cerrado y abierto al mismo tiempo. Es conveniente considerar al conjunto vacío también como abierto y cerrado. Todo conjunto

finito de puntos sobre la recta es cerrado, ya que no tiene puntos límite. El conjunto formado por los puntos

$$0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots, \frac{1}{n}, \dots$$

es cerrado; este conjunto tiene un único punto límite $x = 0$ que pertenece a él.

Nuestra tarea consiste ahora en examinar la estructura de un conjunto cerrado o abierto cualquiera. Para ello necesitamos una serie de hechos auxiliares que admitimos sin demostración.

1. La intersección de un número cualquiera de cerrados es un conjunto cerrado.

2. La unión de un número cualquiera de abiertos es un conjunto abierto.

3. Si un conjunto cerrado está acotado superiormente, contiene a su extremo superior. Análogamente, si un conjunto cerrado está acotado inferiormente, contiene a su extremo inferior.

Sea E un conjunto arbitrario de puntos de la recta. El *complemento* de E , denotado por CE , se define como el conjunto de puntos de la recta que no pertenece a E . Está claro que, si x es un punto exterior de E , es un punto interior de CE , y viceversa.

4. Si un conjunto F es cerrado, entonces su complemento es abierto, y viceversa.

La proposición 4 muestra que existe una relación muy estrecha entre conjuntos abiertos y cerrados: son complemento unos de otros. Debido a esto es suficiente estudiar uno u otro tipo de conjuntos. El conocimiento de las propiedades de uno de los tipos nos permite en el acto descubrir las propiedades del otro. Por ejemplo, un conjunto abierto se obtiene suprimiendo de la recta un conjunto cerrado.

Procedamos ahora al estudio de los conjuntos cerrados. Damos una definición. Sea F un conjunto cerrado. Un intervalo abierto (a, b) se llama *intervalo adyacente* a F si ninguno de sus puntos pertenece a F , mientras que a y b sí pertenecen a F . Entre los intervalos adyacentes también contamos los intervalos impropios (a, ∞) y $(-\infty, b)$, supuesto que a o b pertenezcan a F pero los intervalos no lo interseccionen. Probaremos que si un punto x no pertenece a un conjunto cerrado F , entonces pertenece a alguno de sus intervalos adyacentes.

Denotamos por F_x la parte de F que queda a la derecha de x .

Como x no pertenece a F , podemos representar F_x como una intersección

$$F_x = F \cdot [x, \infty).$$

F y $[x, \infty)$ son ambos cerrados. Por tanto, según la proposición 1, F_x es cerrado. Si F_x es vacío, ningún punto del semiintervalo $[x, \infty)$ pertenece a F . Supongamos que F_x no es vacío. Puesto que este conjunto está contenido en el semiintervalo $[x, \infty)$, está acotado inferiormente. Sea b su extremo inferior. Por la proposición 3, $b \in F_x$ y por tanto $b \in F$. Además, puesto que b es el extremo inferior de F , el semiintervalo $[x, b)$ que queda a la izquierda de b , no contiene puntos de F_x , y por tanto tampoco de F . Hemos construido así un semiintervalo que no contiene puntos de F y tal que, o bien $b = \infty$, o bien b pertenece a F . Análogamente, podemos construir un semiintervalo $(a, x]$ que no contenga puntos de F y tal que $a = -\infty$ o $a \in F$. Es evidente que el intervalo abierto (a, b) contiene a x y es un intervalo adyacente a F . Es fácil comprobar que si (a_1, b_1) y (a_2, b_2) son dos intervalos adyacentes a F , o bien coinciden o son disjuntos.

De lo que precede se deduce que todo conjunto cerrado sobre la recta se obtiene suprimiendo de la recta un cierto número de intervalos abiertos, a saber, los intervalos adyacentes a él. Puesto que todo intervalo abierto contiene al menos un punto racional y los puntos racionales de la recta forman un conjunto numerable, el número de intervalos adyacentes ha de ser, a lo más, numerable. Llegamos por tanto a un notable resultado: todo conjunto cerrado de la recta se obtiene suprimiendo de la recta a lo más un conjunto numerable de intervalos abiertos y disjuntos.

Por la proposición 4 se sigue inmediatamente que todo conjunto abierto es la reunión de, a lo más, un conjunto numerable de intervalos abiertos.

Veremos en el siguiente ejemplo que los conjuntos cerrados pueden tener una estructura muy complicada.

El conjunto perfecto de Cantor

Vamos a construir un conjunto cerrado particular que tiene un cierto número de propiedades notables. Primeramente suprimimos de la recta los intervalos impropios $(-\infty, 0)$ y $(1, \infty)$. Después de esta operación queda el intervalo $[0, 1]$. A continuación suprimimos de éste el intervalo abierto $(1/3, 2/3)$, es decir el tercio intermedio. Este proceso consistente en suprimir el tercio intermedio en los intervalos que van quedando se puede continuar indefinidamente. El con-

junto de puntos que queda después de suprimir todos los intervalos abiertos de la forma descrita se llama conjunto perfecto de Cantor; lo denotaremos por la letra P .

Investiguemos algunas propiedades de este conjunto: P es cerrado, ya que se forma suprimiendo en la recta una cierta colección de intervalos abiertos y disjuntos; P es no vacío, pues por lo menos contiene los puntos extremos de los intervalos suprimidos.

Un conjunto cerrado F se llama *perfecto* si no tiene puntos aislados, es decir, si todos sus puntos son puntos límite. Probemos que P es perfecto. Para ello supongamos que x es un punto aislado de P ; entonces debería ser un extremo común a dos intervalos adyacentes. Pero, por construcción, los intervalos adyacentes a P no tienen extremos comunes.

El conjunto P no contiene ningún intervalo abierto. Para probarlo supongamos que un cierto intervalo abierto δ está contenido en P . Entonces estará contenido en uno de los intervalos obtenidos en el n -ésimo paso de la construcción de P . Ahora bien, esto es imposible, ya que la longitud de los intervalos tiende a cero cuando $n \rightarrow \infty$.

Se puede demostrar que P tiene el número cardinal del continuo. De esto se sigue en particular que P contiene otros puntos además de los puntos extremos de los intervalos adyacentes, ya que éstos forman un conjunto numerable.

En las más diversas ramas de las matemáticas aparecen constantemente distintos tipos de conjuntos de puntos y el conocimiento de sus propiedades es absolutamente indispensable en la resolución de muchos problemas matemáticos. La teoría de conjuntos de puntos es de especial importancia en análisis matemático y en topología.

Vamos a dar ahora nuevos ejemplos de conjuntos de puntos que aparecen en el análisis clásico. Sea $f(x)$ una función continua definida en el intervalo $[a, b]$. Fijamos un número a y consideramos el conjunto de los puntos x tales que $f(x) \geq a$. Es fácil demostrar que este conjunto es un cerrado de $[a, b]$. Análogamente el conjunto de los puntos x tales que $f(x) > a$ es un conjunto abierto $G \subset$ de $[a, b]$. Si $f_1(x), f_2(x), \dots, f_n(x), \dots$ es una sucesión de funciones definidas sobre $[a, b]$, el conjunto de los puntos x en los que la sucesión converge no es un conjunto arbitrario, sino que pertenece a una clase bien definida.

La disciplina matemática cuyo objeto es el estudio de la estructura de los conjuntos de puntos recibe el nombre de *teoría descriptiva de conjuntos*.

En el desarrollo de la teoría descriptiva de conjuntos han hecho

muy importantes aportaciones los matemáticos soviéticos N. N. Luzin y sus discípulos P. S. Aleksandrov, M. Ja. Suslin, A. N. Kolmogorov, M. A. Laurentiev, P. S. Novikov, L. V. Keldys, A. A. Liapunov y otros.

Las investigaciones de N. N. Luzin y sus discípulos han demostrado que las relaciones entre la teoría descriptiva de conjuntos y la lógica formal son muy fuertes. Las dificultades que surgen en un cierto número de problemas de la teoría de conjuntos (en particular, problemas de determinación de la cardinalidad de un conjunto) son de naturaleza lógica. Por otra parte, los métodos de la lógica matemática permiten penetrar más profundamente en algunos problemas de la teoría descriptiva de conjuntos.

§5. *Medida de conjuntos*

El concepto de medida de un conjunto es una fuerte generalización del concepto de longitud de un intervalo. En el caso más simple (el que consideraremos aquí) la tarea consiste en dar una definición de longitud no sólo para intervalos, sino para conjuntos más complicados.

Convenimos que la unidad de medida es el intervalo $[0, 1]$. Entonces la longitud de un intervalo arbitrario $[a, b]$ es obviamente $b - a$. Análogamente, si tenemos dos intervalos disjuntos $[a_1, b_1]$, $[a_2, b_2]$, parece natural tomar como longitud del conjunto E formado por los dos intervalos el número $(b_1 - a_1) + (b_2 - a_2)$. Sin embargo, está menos claro qué debemos entender por longitud de un conjunto más complicado; por ejemplo, ¿cuál es la longitud del conjunto de Cantor del §4 de este capítulo? En conclusión, el concepto de longitud de un conjunto de la recta requiere una definición matemática rigurosa.

El problema de la definición de longitud de conjuntos o, como se dice ahora, de la medida de conjuntos, es muy importante, porque encierra una importancia capital en la generalización del concepto de integral. El concepto de medida de un conjunto tiene aplicaciones a otros problemas en teoría de funciones, probabilidades, topología, análisis funcional, etc.

La definición de medida de un conjunto que describiremos a continuación se debe al matemático francés H. Lebesgue y es la base de la definición de la integral que lleva su nombre.

Medida de conjuntos abiertos y cerrados

Comenzamos con la definición de medida para conjuntos abiertos y cerrados. Como dijimos en §4, todo conjunto abierto es la suma finita o numerable de intervalos abiertos disjuntos dos a dos.

La medida de un conjunto abierto se define como la suma de las longitudes de los intervalos abiertos que lo forman.

Así, si

$$G = \sum (a_i, b_i)$$

y los intervalos (a_i, b_i) son disjuntos dos a dos, entonces la medida de G es igual $\sum (b_i - a_i)$. Denotando la medida de un conjunto E por μE podemos escribir

$$\mu G = \sum (b_i - a_i)$$

En particular, la medida de un intervalo abierto coincide con su longitud.

$$\mu(a, b) = b - a.$$

Todo conjunto cerrado F contenido en $[a, b]$ y tal que los puntos extremos de éste pertenecen a F se obtiene a partir de $[a, b]$ suprimiendo de él un cierto conjunto abierto G . La *medida del conjunto cerrado* $F \subseteq [a, b]$, donde $a \in F$, $b \in F$ se define como la diferencia entre la longitud de $[a, b]$ y la medida del conjunto abierto G complementario de F (relativo a $[a, b]$).

De esta forma

$$\mu F = (a - b) - \mu G \quad (2)$$

No es difícil verificar que, de acuerdo con esta definición, la medida de un intervalo arbitrario es igual a su longitud

$$\mu[a, b] = b - a,$$

y que la medida de un conjunto formado por un número finito de puntos es 0.

Definición general de medida

Con el fin de dar una definición de medida más general, debemos hacer uso de un concepto auxiliar. Sea E un conjunto contenido en el intervalo $[a, b]$. Consideramos todas las posibles *envolturas* de E ,

es decir, todos los conjuntos abiertos $V(E)$ que contienen a E . La medida de cada conjunto $V(E)$ está ya definida. La colección de medidas de todos los conjuntos $V(E)$ es un cierto conjunto de números positivos. Este conjunto está acotado inferiormente (por ejemplo, por 0) y tiene por tanto un extremo inferior que denotamos por $\mu_e E$. El número $\mu_e E$ se llama *medida exterior* de E .

Sea $\mu_e E$ la medida exterior de un conjunto E y $\mu_e CE$ la medida exterior de su complemento relativo a $[a, b]$.

Si la relación

$$\mu_e E + \mu_e CE = b - a \quad (3)$$

se cumple, entonces se dice que E es *medible* y el número $\mu_e E$ es su *medida*: $\mu E = \mu_e E$; si la relación (3) no se cumple, decimos que E no es medible; un conjunto no medible no tiene medida.

Señalemos que siempre

$$\mu_e E + \mu_e CE \geq b - a \quad (4)$$

La medida de los conjuntos más simples (por ejemplo, intervalos abiertos o cerrados) tiene propiedades notables. Mencionamos las más importantes.

1. Si los conjuntos E y E_1 son medibles y $E_1 \subseteq E$, entonces

$$\mu E_1 \leq \mu E;$$

es decir, la medida de una parte de E no debe exceder de la de todo el conjunto E .

2. Si E_1 y E_2 son medibles, entonces el conjunto $E = E_1 + E_2$ es medible y

$$\mu(E_1 + E_2) \leq \mu E_1 + \mu E_2;$$

es decir, la medida de la suma no excede de la suma de las medidas de los sumandos.

3. Si los conjuntos $E_i (i = 1, 2, \dots)$ son medibles y disjuntos dos a dos, $E_i E_j = \emptyset (i \neq j)$, entonces su suma $E = \sum E_i$ es medible y

$$\mu \left(\sum E_i \right) = \sum \mu E_i;$$

es decir, la medida de una suma finita o numerable de conjuntos disjuntos dos a dos es igual a la suma de las medidas de los sumandos.

Esta propiedad de la medida es llamada *aditividad total*.

4. La medida de un conjunto E no cambia si éste se desplaza como un sólido rígido.

Sería deseable que las propiedades de la longitud se conservaran en el concepto más general de medida de conjuntos. Pero se puede probar rigurosamente que esto es *imposible* si la medida se refiere a conjuntos de puntos arbitrarios. En consecuencia, en el sentido de esta definición existen conjuntos que tienen medida (que son medibles) y otros que no tienen medida (que son no medibles). Por otra parte, la clase de conjuntos medibles es tan amplia, que ello no conduce a ninguna desventaja substancial. De hecho, la construcción de un ejemplo de conjunto no medible es bastante difícil.

Presentemos ahora algunos ejemplos de conjuntos medibles.

Ejemplo 1. La medida del conjunto perfecto de Cantor P (ver §4).

En la construcción del conjunto P a partir del intervalo $[0, 1]$ hemos comenzado eliminando un intervalo adyacente de longitud $1/3$, después dos intervalos adyacentes de longitud $1/9$, luego 4 de longitud $1/27$, etc. De forma general, en el paso n -ésimo hemos suprimido 2^{n-1} intervalos adyacentes de longitud $1/3^n$. Así, pues, la longitud de todos los intervalos suprimidos es

$$S = \frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots + \frac{2^{n-1}}{3^n} + \cdots.$$

Los términos de esta serie forman una progresión geométrica de razón $2/3$ cuyo primer término vale $1/3$. Por tanto la suma S de la serie es

$$\frac{\frac{1}{3}}{1 - \frac{2}{3}} = 1.$$

Así, pues, la suma de las longitudes de los intervalos adyacentes al conjunto de Cantor es 1. Dicho de otra forma, la medida del conjunto abierto G complementario de P es 1. Por tanto la medida de P es

$$\mu P = 1 - \mu G = 1 - 1 = 0.$$

Este ejemplo muestra que existen conjuntos con la cardinalidad del continuo cuya medida es 0.

Ejemplo 2. Medida del conjunto R de todos los puntos racionales del intervalo $[1, 0]$.

Mostremos primeramente que $\mu_e R = 0$. En §2 hemos visto que R es numerable. Ordenamos los puntos de R en una sucesión

$$r_1, r_2, \dots, r_n, \dots$$

Ahora, para un $\epsilon > 0$ encerramos el punto r_n en un intervalo abierto δ_n de longitud $\epsilon/2^n$. La suma $\delta = \sum \delta_n$ es un conjunto abierto que contiene a R . Los intervalos δ_n pueden tener intersección no vacía, por lo cual

$$\mu(\delta) = \mu\left(\sum \delta_n\right) \leq \sum \mu\delta_n = \sum \frac{\epsilon}{2^n} = \epsilon.$$

Puesto que ϵ puede ser arbitrariamente pequeño, tenemos $\mu_e R = 0$.

Además, por (2)

$$\mu_e R + \mu_e CR \geq 1,$$

es decir, $\mu_e CR \geq 1$. Pero, como CR está contenido en $[0, 1]$, tenemos que $\mu_e CR \leq 1$. Por consiguiente

$$\mu_e R + \mu_e CR = 1,$$

y ²

$$\mu R = 0, \quad \mu CR = 1. \quad (5)$$

Este ejemplo muestra que un conjunto puede ser denso en un intervalo y sin embargo tener medida 0.

Los conjuntos de medida nula no juegan papel alguno en muchos problemas de la teoría de funciones y pueden ser despreciados. Por ejemplo, una función $f(x)$ es integrable en el sentido de Riemann (integrable Riemann) si, y sólo si, está acotada y el conjunto de los puntos de discontinuidad tiene medida nula. Podríamos añadir un considerable número de ejemplos de este tipo.

Funciones medibles

Procedamos ahora a exponer una de las aplicaciones más brillantes del concepto de medida de un conjunto, a saber, la descripción de la clase de funciones con las que opera actualmente el análisis matemático y la teoría de funciones. El planteamiento preciso es

² El mismo razonamiento muestra que todo conjunto numerable de puntos de la recta tiene medida cero.

como sigue. Si una sucesión $\{f_n(x)\}$ de funciones definidas en un cierto conjunto E converge en cada punto de E excepto en los puntos de un conjunto de medida nula, se dice que la sucesión $\{f_n(x)\}$ converge *en casi todo punto*.

¿Qué funciones se obtienen al aplicar repetidamente las operaciones algebraicas y la operación de paso al límite de una sucesión de funciones convergentes en casi todo punto?

Para contestar a estas preguntas necesitamos algunos conceptos nuevos.

Sea $f(x)$ una función definida en un conjunto E y α un número real arbitrario. Denotamos por

$$E[f(x) > \alpha]$$

el conjunto de los puntos de E tales que $f(x) > \alpha$. Por ejemplo, si $f(x)$ está definido en $[0, 1]$ y $f(x) = x$ sobre este intervalo, el conjunto $E[f(x) > \alpha]$ es igual a $[0, 1]$, para $\alpha < 0$; a $(\alpha, 1]$, para $0 \leq \alpha < 1$, y vacío, para $\alpha \geq 1$.

Una función $f(x)$ definida en un conjunto E se llama medible si E es medible y si el conjunto $E[f(x) > \alpha]$ es medible para cualquier número real α .

Se puede demostrar que cualquier función continua definida en un intervalo es medible. Existen, sin embargo, entre las funciones medibles otras muchas que no son continuas; por ejemplo la función de Direchlet, que es igual a 1 en los puntos irracionales de $[0, 1]$ y a 0 en el resto de los puntos del intervalo.

Mencionemos a continuación, sin demostración, algunas propiedades de las funciones medibles.

1. Si $f(x)$ y $\phi(x)$ son funciones medibles definidas en el mismo conjunto E , entonces las funciones

$$f + \phi, f - \phi, f \cdot \phi \quad \text{y} \quad \frac{f}{\phi}$$

son también medibles (la última si $\phi \neq 0$).

Esta propiedad demuestra que las funciones que se obtienen de aplicar las operaciones algebraicas a funciones medibles son también medibles.

2. Si una sucesión de funciones medibles $\{f_n(x)\}$ definidas en un conjunto E converge en casi todo punto a una función $f(x)$, entonces esta función es también medible.

Esta propiedad nos dice que la operación de paso al límite de una sucesión convergente en casi todo punto de funciones medibles vuelve a dar una función medible.

Estas propiedades de las funciones medibles fueron establecidas por Lebesgue. Un estudio más profundo de las funciones medibles ha sido llevado a cabo por los matemáticos soviéticos D. F. Egorov y N. N. Luzin. En particular N. N. Luzin ha probado que toda función medible definida en un intervalo puede hacerse continua cambiando sus valores en un conjunto de medida arbitrariamente pequeña.

Este resultado clásico de N. N. Luzin y las propiedades enumeradas arriba bastan para probar que las funciones medibles son justamente las funciones por las que nos preguntábamos al comienzo de esta subsección.

Las funciones medibles son también de gran importancia en la teoría de integración: el concepto de integral se puede generalizar de forma que toda función medible y acotada sea integrable. Daremos una exposición detallada de estos problemas en la próxima sección.

§6. *La integral de Lebesgue*

Llegamos al tema central de este capítulo: la definición de la integral de Lebesgue y el estudio de sus propiedades más importantes.

Para comprender el principio fundamental de esta integral, consideremos el siguiente ejemplo. Imaginemos una serie de monedas de distinto valor y queremos hallar cuánto dinero suman en total. Podemos hacerlo de dos maneras. Colocamos las monedas en fila y sumamos el valor de cada una al valor de las precedentes. También podemos seguir un proceso diferente: colocamos las monedas en montones de forma que las monedas de cada montón sean del mismo valor; contamos el número de monedas de cada montón, multiplicamos este número por el valor de la moneda correspondiente, y finalmente sumamos los números así obtenidos. El primer método para contar el dinero corresponde al proceso de integración de Riemann, y el segundo, al proceso de integración de Lebesgue.

Pasando de monedas a funciones, digamos que para el cálculo de la integral de Riemann el dominio de definición de la función (el eje de abscisas; figura 2 a) se divide en pequeñas partes, mientras que para el cálculo de la integral de Lebesgue es el dominio de valores de la función (el eje de ordenadas; figura 2 b) lo que se divide. Este último principio fue aplicado en la práctica mucho antes de Lebesgue,

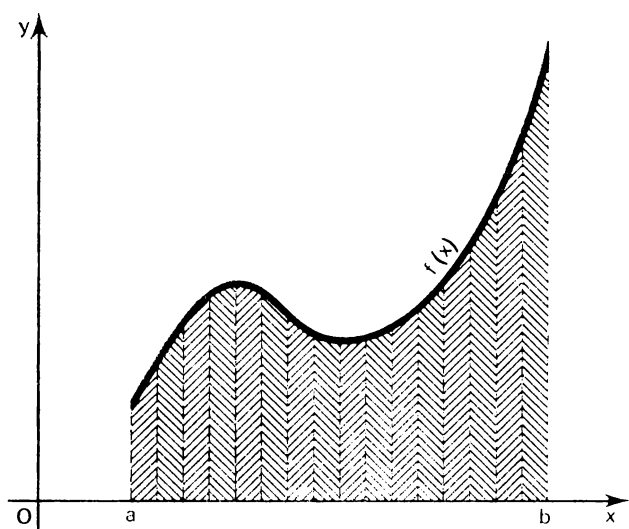


FIG. 2 a

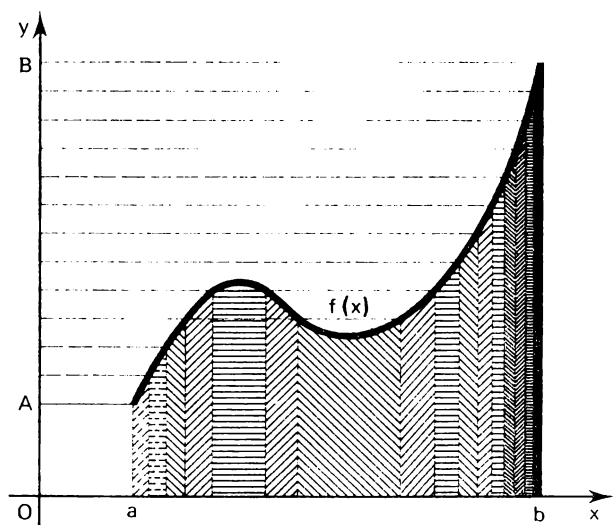


FIG. 2 b

en el cálculo de integrales de funciones de carácter oscilatorio; sin embargo fue Lebesgue quien lo desarrolló con toda generalidad y le dio una fundamentación rigurosa mediante la teoría de la medida.

Examinemos como están relacionadas la teoría de la medida y la integral de Lebesgue. Sea E un conjunto arbitrario medible contenido en el intervalo $[a, b]$. Construimos una función $\phi(x)$ que es igual a 1 si x pertenece a E , y a cero si no pertenece. En otros términos, escribimos

$$\phi(x) = \begin{cases} 1 & x \in E. \\ 0 & x \notin E. \end{cases}$$

La función $\phi(x)$ es llamada habitualmente función característica de E . Consideremos ahora la integral

$$I = \int_a^b \phi(x) dx.$$

Estamos acostumbrados a considerar la integral como el área de la figura D acotada por el eje de abscisas, las rectas $x = a$, $x = b$ y la curva $y = \phi(x)$ (ver capítulo 2). Puesto que en nuestro caso la «altura» de D es distinta de 0 en los puntos $x \in E$ solamente, y en éstos vale 1, el área (por la fórmula del área: longitud por anchura) es igual numéricamente a la longitud (medida) de E . Así, pues, I debe ser igual a la medida de E

$$I = \mu E. \quad (6)$$

La integral de Lebesgue de la función $\phi(x)$ se define precisamente así.

El lector debe observar que la ecuación (6) es la *definición* de la integral $\int_a^b \phi(x) dx$ en el sentido de Lebesgue. Puede suceder que la integral I no exista en el sentido del capítulo 2, es decir, como límite de una suma. Pero, aunque esto ocurra, la integral I siempre existe en el sentido de Lebesgue y vale μE .

Como ejemplo calculamos la integral de la función de Dirichlet $\Phi(x)$, igual a 0 en los puntos racionales y a 1 en los puntos irracionales del intervalo $[0, 1]$. Puesto que la medida del conjunto de los puntos irracionales del intervalo $[0, 1]$ es 1, por (5), la integral de Lebesgue

$$\int_0^1 \Phi(x) dx$$

es igual a 1. Es fácil comprobar que la integral de Riemann de esta función no existe.

Una proposición auxiliar

Supongamos ahora que $f(x)$ es una función arbitraria, acotada y medible. Vamos a demostrar que una función así puede ser representada, con el grado de aproximación que se desee, por una combinación lineal de funciones características de conjuntos. Para ver esto dividamos, en el eje de las ordenadas, el intervalo entre el extremo superior A y el extremo inferior B de los valores de la función por medio de los puntos $y = A, y_1, \dots, y_n = B$ en intervalos de longitud menor que ϵ , donde ϵ es un número positivo previamente fijado. Ahora, si en el punto $x \in [a, b]$,

$$y_i \leq f(x) < y_{i+1} \quad (i = 0, 1, 2, \dots, n-1),$$

entonces ponemos en el punto x

$$\phi(x) = y_i,$$

y si en x

$$f(x) = y_n = B,$$

entonces ponemos

$$\phi(x) = y_n.$$

La construcción de la función $\phi(x)$ se muestra en la figura 3.

Por construcción de $\phi(x)$ tenemos para cada punto de $[a, b]$

$$|f(x) - \phi(x)| < \epsilon.$$

Además, puesto que $\phi(x)$ toma solamente un número finito de valores $y_0, y_1, \dots, y_n$, puede ser escrita en la forma

$$\phi(x) = y_0 \cdot \phi_0(x) + y_1 \cdot \phi_1(x) + \dots + y_n \cdot \phi_n(x) \quad (7)$$

donde $\phi_i(x)$ es la función característica del conjunto de puntos tales que $\phi(x) = y_i$, es decir, $y_i \leq f(x) < y_{i+1}$ (en cada punto $x \in [a, b]$ solamente un sumando de la derecha de (7) es distinto de cero). De esta forma queda probada la proposición.

Definición de la integral de Lebesgue

Procedamos ahora a la definición de la integral de Lebesgue para funciones arbitrarias acotadas y medibles. Como $\phi(x)$ difiere muy

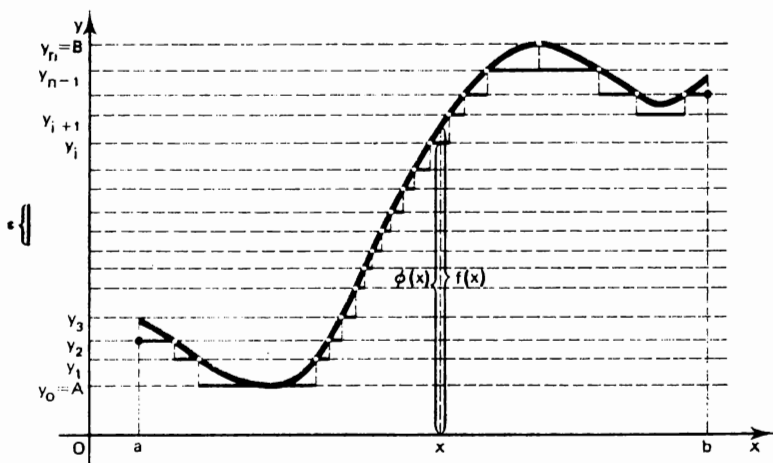


FIG. 3

poco de $f(x)$, podemos tomar como aproximación del valor de la integral de $f(x)$ la de $\phi(x)$. Pero si tenemos en cuenta que las funciones $\phi_i(x)$ son funciones características de conjuntos y usamos formalmente las reglas ordinarias de cálculo de integrales, tenemos

$$\begin{aligned} \int_a^b \phi(x) dx &= \int_a^b \{y_0\phi_0(x) + y_1\phi_1(x) + \cdots + y_n\phi_n(x)\} dx = \\ &= y_0 \int_a^b \phi_0(x) dx + y_1 \int_a^b \phi_1(x) dx + \cdots + y_n \int_a^b \phi_n(x) dx = \\ &= y_0\mu e_0 + y_1\mu e_1 + \cdots + y_n\mu e_n, \end{aligned}$$

donde μe_i es la medida del conjunto e_i de los x tales que

$$y_i \leq f(x) < y_{i+1}$$

Así, un valor aproximado del valor de la integral de Lebesgue de $f(x)$ es la «suma integral de Lebesgue»

$$S = y_0\mu e_0 + y_1\mu e_1 + \cdots + y_n\mu e_n.$$

De acuerdo con esto, la integral de Lebesgue se define como el límite de la suma integral de Lebesgue S cuando

$$\max |y_{i+1} - y_i| \rightarrow 0,$$

que corresponde a la convergencia uniforme de $\phi(x)$ a $f(x)$.

Se puede demostrar que la suma integral de Lebesgue tiene límite para toda función medible acotada, es decir, toda función acotada y medible es integrable en el sentido de Lebesgue. La integral de Lebesgue se puede extender a cierta clase de funciones medibles no acotadas, pero aquí no trataremos de ello.

Propiedades de la integral de Lebesgue

La integral de Lebesgue tiene todas las propiedades de la integral ordinaria, a saber, la integral de la suma es igual a la suma de las integrales, una constante permuta con el signo integral, etc. Sin embargo la integral de Lebesgue tiene una importante propiedad que la integral ordinaria no tiene: si las funciones $f_n(x)$ son medibles y están uniformemente acotadas

$$|f_n(x)| < K$$

para todo n y todo x en $[a, b]$ y si la sucesión $\{f_n(x)\}$ converge en casi todo punto a $f(x)$, entonces

$$\int_a^b f_n(x) dx \rightarrow \int_a^b f(x) dx.$$

En otros términos, la integral de Lebesgue permite sin restricción el paso al límite. De hecho esta propiedad hace de la integral de Lebesgue una herramienta muy conveniente y a veces indispensable en muchas investigaciones. En particular, la teoría de la integral de Lebesgue es indispensable en la teoría de series trigonométricas, en la de espacios funcionales (ver el capítulo 18) y en otras ramas de las matemáticas.

Demos un ejemplo. Sea $f(x)$ una función periódica de período 2π y

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \operatorname{sen} nx)$$

su serie de Fourier. Si, por ejemplo, $f(x)$ es continua, es fácil demostrar que

$$\frac{1}{\pi} \int_0^{2\pi} f^2(x) dx = \frac{a_0^2}{2} + \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \quad (8)$$

Esta identidad se conoce con el nombre de igualdad de Parseval. Se plantea ahora la siguiente pregunta: ¿para qué tipo de funciones es válida la igualdad de Parseval (8)? La respuesta es: la igualdad de Parseval es válida si, y sólo si, $f(x)$ es medible sobre $[0, 2\pi]$ y $f^2(x)$ es integrable Lebesgue en el mismo intervalo.

§1. El alcance del álgebra lineal y sus métodos

Funciones lineales y matrices

Entre las funciones de una sola variable, la más simple, con mucho, es la llamada *función lineal* $l(x) = ax + b$. Su gráfica es, desde luego, la más simple de las curvas: la línea recta.

Al mismo tiempo, la función lineal es una de las más importantes. Esto se debe a que cualquier «trozo» pequeño de una curva «lisa» se puede equiparar a una línea recta y a que cuanto más pequeño es el «trozo», mayor es el parecido. En el lenguaje de la teoría de funciones, esto significa que toda función «lisa» (continuamente diferenciable) es, para un pequeño cambio de la variable independiente, muy próxima a una función lineal. La función lineal se caracteriza por el hecho de que su incremento es proporcional al incremento de la variable independiente. Efectivamente: $\Delta l(x) = l(x_0 + \Delta x) - l(x_0) = a(x_0 + \Delta x) + b - (ax_0 + b) = a\Delta x$. Inversamente, si $\Delta l(x) = a\Delta x$, entonces $l(x) - l(x_0) = a(x - x_0)$ y $l(x) = ax + l(x_0) - ax_0 = ax + b$, donde $b = l(x_0) - ax_0$. Pero por el cálculo diferencial sabemos que en el incremento de una función diferencial arbitraria podemos aislar de forma natural la parte principal, llamada diferencial de la función, la cual es proporcional al incremento de la variable independiente, y que el incremento de la función difiere de su di-

ferencial en un infinitésimo de orden mayor que el de la variable independiente. Así, para un cambio infinitamente pequeño de la variable independiente, una función diferenciable se aproxima a una función lineal con un error menor que un infinitésimo de orden mayor que el de la variable independiente.

La situación con funciones de varias variables es similar.

Una *función lineal de varias variables* es una función de la forma $a_1x_1 + a_2x_2 + \cdots + a_nx_n + b$. Si $b = 0$, la función se llama homogénea. Una función lineal de varias variables se caracteriza por las dos propiedades siguientes:

1. El incremento de una función lineal, calculado cuando sólo una de las variables independientes recibe un incremento, permaneciendo fijas las demás, es proporcional al incremento de dicha variable.

2. El incremento de una función lineal, calculado cuando se incrementan todas las variables independientes, es igual a la suma algebraica de los incrementos obtenidos cuando se incrementa cada variable por separado.

La función lineal de varias variables juega el mismo papel entre las funciones de ese número de variables que la función lineal de una variable entre las funciones de una sola variable. Pues una función «lisa» (es decir, una función con derivadas parciales continuas respecto a cada una de las variables) está muy próxima a una función lineal para pequeñas variaciones de la variable independiente. De hecho, el incremento de una función así, $w = f(x_1, x_2, \dots, x_n)$ es igual, salvo infinitésimos de orden superior, a la diferencial total $(\partial f / \partial x_1)dx_1 + \cdots + (\partial f / \partial x_n)dx_n$, que es una función lineal homogénea de los incrementos $dx_1, \dots, dx_n$ de las variables independientes. De aquí que la propia función w , que es igual a la suma de su valor inicial más el incremento, pueda ser expresada en función de sus variables independientes, para pequeños cambios de éstas, en forma de función lineal no homogénea con un error menor que un infinitésimo de orden superior.

Los problemas cuya solución requiere la investigación sobre funciones de varias variables surgen en conexión con el estudio de la dependencia de una cantidad respecto de otras varias. Un problema se llama *lineal* si la dependencia considerada es lineal. Por las propiedades de las funciones lineales citadas arriba, un problema lineal se caracteriza por:

1. *La propiedad de proporcionalidad.* La repercusión de cada factor por separado es proporcional a su valor.

El hecho de que una función «lisa» pueda ser reemplazada en primera aproximación por otra lineal refleja un principio general conocido: cualquier problema en el que intervenga la variación de una cantidad como consecuencia de la acción de varios factores puede ser interpretado, en primera aproximación y para pequeñas acciones, como un problema lineal, es decir, como si tuviera las propiedades de linealidad e independencia. Ocurre a menudo que este punto de vista conduce a resultados adecuados en la práctica (la teoría clásica de elasticidad, teoría de pequeñas oscilaciones, etc.)

Una función lineal tiene propiedades tan simples que no requiere un estudio especial. Otra cosa son las funciones lineales de varias variables, donde la presencia de más de una variable introduce algunas características especiales. La situación se complica mucho más cuando pasamos de una función de varias variables $x_1, x_2, \dots, x_n$ a un conjunto de varias funciones $y_1, y_2, \dots, y_n$ de las mismas variables. Como «primera aproximación» aparece un sistema de funciones lineales:

$$\begin{aligned} y_1 &= a_{11}x_1 + \cdots + a_{1n}x_n + b_1, \\ y_2 &= a_{21}x_1 + \cdots + a_{2n}x_n + b_2, \\ &\vdots \\ y_m &= a_{m1}x_1 + \cdots + a_{mn}x_n + b_m. \end{aligned}$$

El estudio de funciones lineales y de sus sistemas también constituye el objeto inicial de una rama del álgebra llamada álgebra lineal.

Casos especiales e importantes de matrices son las formadas por una sola columna, llamadas simplemente columnas o matrices columna, las que constan solamente de una fila, llamadas filas o matrices fila, y finalmente las matrices cuadradas, es decir, las que tienen el mismo número de filas que de columnas. El número de filas (o columnas) de una matriz cuadrada se llama *orden* de la matriz. La «matriz» (a) formada por un único número se identifica con el propio número.

Las operaciones con matrices se definen en relación con las operaciones entre formas lineales.

Supongamos dos sistemas de formas lineales

[illegible]

y

$$\begin{aligned} z_1 &= b_{11}x_1 + \cdots + b_{1n}x_n, \\ &\dots\dots\dots \\ z_m &= b_{m1}x_1 + \cdots + b_{mn}x_n. \end{aligned}$$

Sumemos estas dos formas:

$$\begin{aligned} y_1 + z_1 &= (a_{11} + b_{11})x_1 + \cdots + (a_{1n} + b_{1n})x_n, \\ &\vdots \\ y_m + z_m &= (a_{m1} + b_{m1})x_1 + \cdots + (a_{mn} + b_{mn})x_n. \end{aligned}$$

Parece natural decir que la matriz del sistema de formas así obtenido

$$\begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$

es la suma de las matrices

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad y \quad \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix}.$$

La matriz de coeficientes

$$\begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{k1} & \cdots & c_{kn} \end{bmatrix}$$

se dice que es el producto de las matrices

$$\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{km} \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix}$$

y se escribe

$$\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{km} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix}$$

Es fácil calcular cómo se hallan los elementos de la matriz producto en función de los elementos de los factores. El elemento c_{ij} es el coeficiente de x_j en la expresión de z_i en función de $x_1, x_2, \dots, x_n$.

Pero si $z_i = a_{i1}y_1 + \cdots + a_{im}y_m$, y por otra parte

$$\begin{aligned} y_1 &= \cdots + b_{1j}x_j + \cdots, \\ &\vdots \\ y_m &= \cdots + b_{mj}x_j + \cdots. \end{aligned}$$

Por tanto,

$$z_i = \cdots + (a_{i1}b_{1j} + \cdots + a_{im}b_{mj})x_j + \cdots,$$

y de aquí

$$c_{ij} = a_{i1}b_{1j} + \cdots + a_{im}b_{mj}.$$

Así, pues, el elemento de la fila i y la columna j del producto de dos matrices es igual a la suma de los productos de los elementos de la fila i del primer factor por los correspondientes elementos de la columna j del segundo factor. Por ejemplo:

$$\begin{pmatrix} 2 & 1 & 3 \\ 3 & 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 1 & 2 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 \cdot 3 + 1 \cdot 1 + 3 \cdot 2 & 2 \cdot 1 + 1 \cdot 2 + 3 \cdot 4 \\ 3 \cdot 3 + 1 \cdot 1 + 1 \cdot 2 & 3 \cdot 1 + 1 \cdot 2 + 1 \cdot 4 \end{pmatrix} = \begin{pmatrix} 13 & 16 \\ 12 & 9 \end{pmatrix}.$$

Enumeremos las propiedades fundamentales de las operaciones con matrices.

1. $A + B = B + A$ (conmutatividad de la adición)
2. $(A + B) + C = A + (B + C)$ (asociatividad de la adición)
3. $c(A + B) = cA + cB$ } (leyes distributivas para la multiplicación por un número. Aquí c, c_1, c_2 son números y no matrices).
- 3'. $(c_1 + c_2)A = c_1A + c_2A$ }
4. $(c_1c_2)A = c_1(c_2A)$ (ley asociativa para la multiplicación por un número)
5. Existe una matriz «nula»

$$O = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

tal que $A + O = A$ para cualquier matriz A .

6. $c \cdot O = 0 \cdot A = O$; inversamente si $cA = O$, entonces $c = 0$ ó $A = O$ (aquí c es un número).

7. Para cada matriz A existe una matriz opuesta $-A$, es decir, tal que $A + (-A) = O$.

8. $(A + B) \cdot C = AC + BC$ }
- 8'. $C(A + B) = CA + CB$ } (leyes distributivas para la adición y multiplicación de matrices).
9. $(AB)C = A(BC)$ (leyes asociativa para la multiplicación).

$$10. (cA)B = A(cB) = c(AB)$$

Estas propiedades se cumplen no sólo para matrices cuadradas, sino también para matrices rectangulares cualesquiera, a condición de que las operaciones estén definidas. Para matrices cuadradas esta condición se cumplen automáticamente.

Todas estas propiedades de las operaciones con matrices son similares a las propiedades de las operaciones con números.

Vamos a señalar ahora dos particularidades de las operaciones con matrices. Primero, la multiplicación no tiene propiedad conmutativa, es decir, AB no es siempre igual a BA . Por ejemplo:

$$\begin{pmatrix} 3 & -2 \\ -1 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = \begin{pmatrix} -3 & 2 \\ 11 & 6 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 3 & -2 \\ -1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 6 \\ 7 & 2 \end{pmatrix}$$

En segundo lugar, el producto de dos números es cero si, y sólo si, uno de los dos factores es cero. Este teorema es fundamental en la teoría de ecuaciones algebraicas. Pero es falso para la multiplicación de matrices. El producto de dos matrices puede ser la matriz nula sin que lo sea ninguno de los dos factores. Por ejemplo:

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Mencionemos todavía otra propiedad de la multiplicación de matrices. La matriz $\bar{A}$ se llama *traspuesta* de A si cada fila de $\bar{A}$ está formada por los elementos de la columna correspondiente de A en el mismo orden. Por ejemplo, para la matriz

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

su traspuesta es la matriz

$$\bar{A} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

La multiplicación se relaciona con la trasposición mediante la siguiente fórmula:

$$\overline{AB} = \bar{B}\bar{A},$$

que se verifica fácilmente mediante la regla de multiplicación de matrices.

La teoría de matrices constituye una parte indispensable del álgebra lineal en tanto en cuanto juega el papel de un método para el planteamiento y resolución de sus problemas.

Analogías geométricas en álgebra lineal

Aparte de la fuente ya descrita de la que surgen ideas y problemas del álgebra lineal, están también las necesidades del análisis matemático y de la geometría, en particular de la geometría analítica, que conducen también al desarrollo de aquella disciplina, enriqueciéndola con importantes ideas y analogías. Es bien sabido que la geometría analítica del plano, y en mayor medida aún la del espacio, hacen uso, en lo que conciernen a la teoría de rectas y planos, de los métodos del álgebra lineal en su forma más simple. Una recta en el

plano viene dada por una ecuación lineal de dos variables que liga las dos coordenadas de un punto arbitrario de la recta. Un plano en el espacio viene dado por una ecuación lineal en tres variables (las tres coordenadas de un punto arbitrario del plano); una recta en el espacio, por dos ecuaciones lineales.

Por supuesto, el uso del concepto de vector confiere a la geometría analítica, y en consecuencia a la teoría de ecuaciones lineales, una gran simplicidad y claridad. Pues bien, el uso del concepto de vector en sentido generalizado ofrece las mismas ventajas al álgebra lineal, sobre todo en la teoría general de sistemas de ecuaciones lineales. El camino de esta generalización es el siguiente. Un vector (en el espacio) se da mediante tres números, a saber, sus tres proyecciones sobre los ejes coordenados. Cada terna de números reales, a su vez, puede ser representada geoméricamente en forma de un vector (en el espacio).

Las operaciones de adición (por la «ley del paralelogramo») y multiplicación por un número están definidas para los vectores. Y lo están en relación con las operaciones similares con fuerzas, velocidades, aceleraciones, y otras cantidades físicas que se pueden representar por vectores.

Si los vectores vienen dados por sus coordenadas (es decir por sus proyecciones sobre los ejes coordenados), entonces las operaciones de adición y multiplicación por un número realizadas con vectores corresponden a las mismas operaciones con las filas (o columnas) formadas con sus coordenadas.

Así, es conveniente interpretar una fila o columna de tres elementos geoméricamente como un vector en el espacio tridimensional. De esta forma, las operaciones básicas con filas o columnas pueden interpretarse como las correspondientes operaciones con vectores del espacio, de modo que el álgebra de filas o columnas de tres elementos no difiere formalmente del álgebra de vectores en el espacio tridimensional. Esta circunstancia hace que sea natural introducir una terminología geométrica en el álgebra lineal.

Una columna (o fila) de n números

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}$$

se considera como un «vector», es decir, como un elemento de un «espacio vectorial n -dimensional». La suma de los vectores

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{y} \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

es el vector

$$\begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \dots \\ x_n + y_n \end{pmatrix};$$

el producto del vector

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

por el número c es el vector

$$\begin{pmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_n \end{pmatrix}$$

El conjunto de todos los vectores (columnas) forman, por definición, el espacio vectorial aritmético n -dimensional.

Junto con el espacio vectorial aritmético n -dimensional podemos introducir el concepto de espacio de puntos n -dimensional, asociando a cada columna una imagen geométrica: un punto. Entonces el espacio vectorial n -dimensional se define de la siguiente manera. A cada par de puntos A y B asociamos el vector AB que va de A a B tomando como coordenadas del vector (sus proyecciones sobre los ejes coordenados), por definición, la diferencia entre las correspondientes

coordenadas de los puntos A y B . Se dice que dos vectores son iguales si sus correspondientes coordenadas son iguales, exactamente igual que en el espacio tridimensional, donde dos vectores se consideran iguales si uno de ellos se obtiene a partir del otro mediante una traslación.

Entre los vectores de un espacio vectorial n -dimensional y los puntos de un espacio de puntos n -dimensional, existe una correspondencia uno a uno que se construye de un modo natural.

El punto

$$\begin{pmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$$

se toma como origen de coordenadas, y a cada punto se le hace corresponder el vector que une el origen de coordenadas con ese punto. Por otra parte, a cada vector se le asocia su punto extremo, suponiendo que su origen coincide con el origen de coordenadas. La introducción del espacio de puntos crea nuevas analogías que permiten «ver» mejor en el espacio n -dimensional.

Sin embargo, en ulteriores generalizaciones (§2) la definición rigurosa de espacio de puntos se hace un poco más complicada, y por eso no haremos uso de este concepto. El lector que quiera hacer uso de las analogías que nacen del estudio de los espacios de puntos deberá visualizar los elementos de un espacio vectorial mediante vectores que parten del origen de coordenadas.

La introducción de una terminología geométrica permite usar en álgebra lineal las analogías basadas en la intuición geométrica que se origina con el estudio de la geometría en el espacio tridimensional.

Por supuesto, las analogías se deben usar con cierta cautela, teniendo en cuenta que los argumentos geométrico-intuitivos pueden comprobarse con una lógica estricta, haciendo uso solamente de las definiciones de los conceptos «geométricos» y de los teoremas rigurosamente demostrados.

Una característica notable de los elementos de un espacio vectorial n -dimensional es la existencia de las operaciones de adición y multiplicación por un número, con propiedades que recuerdan a las operaciones con números. Como ya hemos mencionado al hablar

de las propiedades de las operaciones con matrices, la adición satisface las leyes conmutativa y asociativa, se cumplen las leyes distributivas (para la multiplicación por un número), la adición tiene un único inverso, y el producto de un vector por un número es el vector nulo si, y sólo si, el vector es el vector nulo o el número es cero.

Sin embargo, no sólo las columnas (o filas) tienen las propiedades referidas. Tales características también son propias del conjunto de matrices de la misma estructura, y de las magnitudes físicas de tipo vectorial: fuerzas, velocidades, aceleraciones, etc. Son incluso propias de objetos matemáticos de naturaleza enteramente distinta; como por ejemplo el conjunto de todos los polinomios de una variable, el conjunto de todas las funciones continuas definidas en un intervalo dado $[a, b]$, el conjunto de todas las soluciones de una ecuación diferencial lineal homogénea, etc.

Esta circunstancia motiva una nueva generalización de espacio vectorial, a saber, la introducción de espacios lineales generales. Los elementos de tales espacios pueden ser objetos matemáticos o físicos arbitrarios para los que las operaciones de adición y multiplicación por un número se definen de forma natural. Un enfoque tan general y abstracto del concepto de espacio lineal no introduce ninguna clase de complicaciones en la teoría, tal y como hemos visto anteriormente: un espacio lineal (y desde luego, un espacio n -dimensional; el significado de esto será aclarado en la próxima sección) no difiere en su estructura y propiedades del espacio lineal aritmético, pero el campo de aplicación es considerablemente mayor gracias a esta generalización, permitiendo la aplicación de los métodos del álgebra lineal a una clase mayor de problemas de la ciencia teórica.

§2. Espacios lineales

Definición de un espacio lineal

Vamos a proceder a definir rigurosamente el espacio lineal.

Un *espacio lineal* es una colección de objetos de naturaleza arbitraria para los cuales tiene sentido la suma y el producto por un número, satisfaciendo los siguientes postulados:

1. $(X + Y) + Z = X + (Y + Z)$.
2. Existe un elemento «nulo», el 0 , tal que $X + 0 = X$ para todo X .
3. Para cada elemento X existe un opuesto $-X$ tal que $X + (-X) = 0$.

4. $X + Y = Y + X$.
5. $1 \cdot X = X$.
6. $c_1(c_2X) = (c_1c_2)X$.
7. $(c_1 + c_2)X = c_1X + c_2X$.
8. $c(X + Y) = cX + cY$.

Aquí, X, Y, Z son elementos del espacio lineal; $1, c_1, c_2, c$ son números.

Estos postulados (llamados también *axiomas del espacio lineal*) son muy naturales y constituyen una exposición formal de aquellas propiedades de la adición y multiplicación por un número que deben necesariamente acompañar a estas operaciones cualquiera que sea el sentido en el que deban ser interpretadas. Las operaciones con cualquier significado —por ejemplo, físico— que satisfagan los postulados 1-8 se consideran siempre como una suma y una multiplicación por un número.

A continuación enunciamos algunas consecuencias de los axiomas:

- a. El elemento nulo 0 es único, es decir, existe un único elemento que satisface el axioma 2.
- b. El elemento opuesto de un elemento dado X es único.
- c. La «sustracción» tiene sentido; es decir, dada una suma y uno de los sumandos, el otro está definido de manera única: si $X + Y = Z$, entonces $Y = Z + (-X)$.
- d. $0 \cdot X = c \cdot 0 = 0$.
- e. Si $cX = 0$, entonces o bien $c = 0$, o bien $X = 0$.
- f. $-X = (-1)X$.

La demostración de estas consecuencias es muy sencilla y se deja como ejercicio. En lo que sigue, los elementos de un espacio lineal serán llamados vectores.

Dependencia e independencia lineal de vectores

Pasamos ahora al importante concepto de dependencia e independencia lineal de vectores.

El vector $c_1X_1 + c_2X_2 + \cdots + c_mX_m$, para cualquier valor numérico de los coeficientes $c_1, c_2, \cdots, c_m$, recibe el nombre de combinación lineal de los vectores $X_1, X_2, \cdots, X_m$. Si entre los vectores $X_1, X_2, \cdots, X_m$ hay al menos uno que es combinación lineal de los restantes, se dice que los vectores $X_1, X_2, \cdots, X_m$ son linealmente dependientes. Pero si ninguno de ellos es combinación lineal de los demás, entonces se dice que son linealmente independientes.

Es fácil ver que para que los vectores sean linealmente independientes es necesario y suficiente que la relación $c_1X_1 + c_2X_2 + \cdots + c_mX_m = 0$ sea cierta sólo si $c_1 = c_2 = \cdots = c_m = 0$.

Para vectores del espacio tridimensional ordinario los conceptos de dependencia e independencia lineal tienen una interpretación geométrica sencilla.

Supongamos dados dos vectores X_1, X_2 . Que los dos vectores son linealmente dependientes quiere decir que uno de ellos es «combinación lineal» del otro, es decir, que sólo difieren en un factor numérico. Esto significa que ambos están contenidos en la misma recta, es decir, que tienen la misma dirección con igual u opuesto sentido.

Inversamente, si dos vectores están contenidos en la misma recta, son linealmente dependientes. En consecuencia, la independencia lineal de dos vectores X_1 y X_2 significa que no pueden estar situados en la misma recta; sus direcciones son esencialmente distintas.

Veamos ahora lo que significa la dependencia e independencia lineal de tres vectores. Supongamos que los vectores X_1, X_2, X_3 , son linealmente dependientes, y como consecuencia de la definición, que el vector X_3 es una combinación lineal de los vectores X_1, X_2 . Entonces X_3 debe estar contenido en el plano que contiene a X_1 y X_2 ; es decir los tres vectores están en el mismo plano. Es fácil ver que si los vectores X_1, X_2, X_3 están en el mismo plano, entonces son linealmente dependientes. Para ello supongamos que X_1, X_2 no están sobre la misma recta; entonces X_3 se puede descomponer con respecto a X_1, X_2 , es decir, se puede representar como combinación lineal de X_1, X_2 . Por otra parte, si X_1, X_2 están en la misma recta, la dependencia lineal es evidente.

Así, pues, la dependencia lineal de tres vectores X_1, X_2, X_3 es equivalente al hecho de que los tres vectores estén en el mismo plano. Y por tanto X_1, X_2, X_3 son linealmente independientes si, y sólo si, no están contenidos en un mismo plano.

Cuatro vectores en el espacio tridimensional son siempre linealmente dependientes. Para verlo supongamos primero que X_1, X_2, X_3 son linealmente dependientes; entonces los vectores X_1, X_2, X_3, X_4 son linealmente dependientes cualquiera que sea X_4 . Si, por el contrario, los vectores X_1, X_2, X_3 son linealmente independientes, entonces no están en el mismo plano y cualquier otro vector X_4 se puede descomponer respecto a X_1, X_2, X_3 , es decir, se puede representar como una combinación lineal de ellos.

Los razonamientos precedentes se pueden generalizar de la siguiente forma.

En el espacio tridimensional los vectores $X_1, X_2, \dots, X_k (k \geq 3)$ son linealmente dependientes si, y sólo si, pertenecen a un espacio (recto o plano) de dimensión menor que k .

En las páginas siguientes daremos una definición rigurosa de subespacio y de dimensión, de forma que también en el caso general la dependencia lineal de los vectores $X_1, X_2, \dots, X_k$ sea equivalente a que pertenezcan a un subespacio de dimensión menor que k ; es decir, el significado geométrico de dependencia lineal será el mismo en el espacio lineal general que en el espacio tridimensional.

El siguiente teorema juega un papel fundamental en la teoría de espacios lineales: si los vectores $X_1, X_2, \dots, X_m$ son combinaciones lineales de los vectores $Y_1, Y_2, \dots, Y_k$ y $m > k$, entonces $X_1, X_2, \dots, X_m$ son linealmente dependientes (teorema sobre la dependencia lineal de combinaciones lineales).

Para $k = 1$ el teorema es obvio. Para $k > 1$ se puede probar fácilmente mediante inducción matemática respecto a k .

Bases y dimensión de un espacio

En el espacio tridimensional, toda terna de vectores que no estén en un plano (es decir, que sean linealmente independientes) forman una *base* del espacio. Esto quiere decir que, si los tres vectores son X_1, X_2, X_3 , cualquier otro vector del espacio se puede escribir como combinación lineal de ellos.

Los espacios lineales generales se pueden dividir en dos tipos.

Puede ocurrir que exista un número arbitrariamente grande de vectores linealmente independientes. Tales espacios se llaman de dimensión infinita y su estudio conduce a una rama del álgebra lineal que es parte importante de una disciplina matemática especial llamada análisis funcional (ver el capítulo 8).

Un espacio lineal se dice que es de dimensión finita si existe una cota finita para el número de vectores linealmente independientes, es decir, un número n tal que existen en el espacio n vectores linealmente independientes, pero no más de n . El número n se llama *dimensión* del espacio.

Así, el espacio de vectores del espacio geométrico tridimensional ordinario es también tridimensional en el sentido que acabamos de indicar. Basta observar que en el espacio ordinario existen muchas ternas de vectores lineales linealmente independientes, pero nunca cuaternas con esta propiedad.

El espacio de vectores columna de n términos es n -dimensional

en el sentido de nuestra definición. Basta ver que existen n vectores linealmente independientes, como por ejemplo:

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix} \cdots, \quad e_n = \begin{pmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}, \quad (2)$$

pero cualquier vector

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}$$

del espacio es combinación lineal de ellos, a saber: $x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$. Por tanto, por el teorema de la dependencia lineal de combinaciones lineales no puede haber más de n vectores linealmente independientes.

Los polinomios de una variable forman un espacio lineal. Basta comprobar que las definiciones usuales de adición y multiplicación por un número satisfacen los axiomas 1-8. Sin embargo, este espacio es de dimensión infinita, ya que los vectores $1, x, \cdots, x^N$ son linealmente independientes para cualquier N . Pero el conjunto de polinomios cuyo grado no excede de N es un espacio de dimensión finita, de dimensión $N + 1$. Pues los vectores $1, x, \cdots, x^N$ son linealmente independientes y son en número $N + 1$. Ahora bien, cualquier polinomio de grado no superior a N es combinación lineal de $1, x, \cdots, x^N$, y por el teorema de la independencia lineal no puede haber más de $N + 1$ polinomios de grado N linealmente independientes.

Introducimos ahora el importante concepto de *base* de un espacio n -dimensional. Una base se define como un conjunto de vectores linealmente independientes tal que todo vector del espacio es combinación lineal de los vectores de dicho conjunto. Así, en el espacio de los vectores columna, el conjunto de vectores (2) es una base. En el espacio de los polinomios de grado $\leq N$, los «vectores» $1, x, \cdots, x^N$ forman una base. En el espacio geométrico tridimen-

sional, cada terna de vectores linealmente independientes juega el papel de una base.

En un espacio lineal n -dimensional, cada conjunto de n vectores linealmente independiente (y la existencia de al menos uno de estos conjuntos está asegurada por la definición de espacio n -dimensional) es una base del espacio. Para probarlo supongamos que $e_1, e_2, \dots, e_n$ son vectores linealmente independientes de un espacio lineal y X es un vector arbitrario del mismo espacio. Entonces los vectores $X, e_1, e_2, \dots, e_n$ son linealmente dependientes (ya que son más de n), es decir, existe un conjunto de números $c, c_1, c_2, \dots, c_n$, no todos iguales a 0, tales que $cX + c_1e_1 + \dots + c_ne_n = 0$. Además $c \neq 0$, pues si $c = 0$, los vectores $e_1, e_2, \dots, e_n$ serían linealmente dependientes. Por tanto, $X = -(c_1/c)e_1 - \dots - (c_n/c)e_n$; es decir, cualquier vector del espacio es combinación lineal de los vectores $e_1, e_2, \dots, e_n$.

Toda base de un espacio lineal n -dimensional consta exactamente de n vectores. Como los vectores de una base son linealmente independientes, su número no puede ser mayor que n . Sea, por otra parte, $e_1, e_2, \dots, e_k$ una base del espacio n -dimensional. Como acabamos de ver $k \leq n$. Pero todo vector del espacio, por definición de base, es combinación lineal de los vectores $e_1, e_2, \dots, e_k$, y, por el teorema sobre la dependencia lineal de combinaciones lineales, no puede haber más de k vectores linealmente independientes, de donde se deduce que la dimensión del espacio n no puede ser mayor que el número de vectores de la base. Así, pues, $k = n$, como queríamos demostrar.

Introducimos ahora el concepto de *coordenadas* de un vector respecto a una base $e_1, e_2, \dots, e_k$. Como hemos visto, cada vector X es combinación lineal de los vectores de la base. Esta representación es además única. Para probarlo supongamos que se puede expresar un cierto vector X en función de la base $e_1, e_2, \dots, e_n$ de dos formas distintas:

$$\begin{aligned} X &= x_1e_1 + x_2e_2 + \dots + x_ne_n, \\ X &= x'_1e_1 + x'_2e_2 + \dots + x'_ne_n. \end{aligned}$$

Entonces $(x_1 - x'_1)e_1 + (x_2 - x'_2)e_2 + \dots + (x_n - x'_n)e_n = 0$, y de la independencia lineal de $e_1, e_2, \dots, e_n$ se sigue que $x_1 = x'_1, \dots, x_n = x'_n$.

Los coeficientes $x_1, x_2, \dots, x_n$ de la descomposición de un vector arbitrario X en función de los vectores de una base se llaman *coordenadas* de X en dicha base. De esta forma, a todo vector, una vez

escogida la base, se le puede asociar la fila (o columna) de sus coordenadas, y viceversa: toda fila (o columna) de n números se puede considerar como el conjunto de coordenadas de un cierto vector.

Las operaciones de adición de vectores y multiplicación de un vector por un número corresponden a las operaciones similares con las filas (o columnas) correspondientes.

Por tanto, todo espacio lineal n -dimensional, independientemente de la naturaleza de sus elementos (pueden ser funciones, matrices, cantidades físicas, etc.) no difiere en nada del correspondiente espacio de filas (o columnas) respecto de estas operaciones. Así, como ya hemos mencionado, la axiomática generalizada que introduce el concepto de espacio lineal no produce complicaciones al considerar un espacio como espacio de filas, sino que extiende considerablemente el campo de aplicación de este concepto.

Una identidad de propiedades entre dos conjuntos de objetos en relación con un sistema de operaciones (o con otro sistema arbitrario de relaciones entre sus elementos) se llama en matemáticas un isomorfismo. La definición exacta de isomorfismo de sistemas algebraicos se dará en el capítulo 19. Usando este término, podemos decir que todos los espacios lineales n -dimensionales, independientemente de la naturaleza de sus elementos, son isomorfos entre sí e isomorfos a un único modelo, a saber, el espacio de filas formadas por n elementos.

Subespacios

Un conjunto de vectores de un espacio lineal n -dimensional R_n , se dice que es un *subespacio* de R si toda combinación lineal de vectores de dicho conjunto también pertenece a él. Evidentemente un *subespacio* de R_n es también un espacio lineal y tiene, por tanto, bases y una dimensión. Es también evidente que la dimensión del subespacio no excede de la dimensión del espacio total y que es igual a ella si, y sólo si, el subespacio coincide con el espacio.

Ejemplos de subespacios del espacio tridimensional de vectores son las rectas y los planos, salvo una traslación, o más precisamente los conjuntos compuestos por todos los vectores contenidos en un plano o en una recta.

Muy frecuentemente hay que estudiar los subespacios «generados» por un sistema de vectores. Estos subespacios se definen como sigue. Supongamos dado un sistema de vectores $X_1, X_2, \dots, X_n$ linealmente dependientes o independientes. Entonces, el conjunto de todas las combinaciones lineales $\{c_1X_1 + c_2X_2 + \dots + c_nX_n\}$ es un subespacio

de R_n que recibe el nombre de subespacio generado por los vectores $X_1, X_2, \dots, X_n$.

La dimensión de este subespacio se llama *rango* del sistema de vectores $X_1, X_2, \dots, X_n$. Es fácil ver que el rango de un sistema de vectores es igual al máximo número de vectores linealmente independientes contenido en el sistema.

El «conjunto» formado solamente por el vector nulo satisface formalmente las condiciones impuestas a un subespacio. Como dimensión de este subespacio se forma el número cero.

Dados dos subespacios de un espacio R_n , podemos formar a partir de ellos de una manera natural otros dos subespacios: su suma vectorial (o unión) y su intersección.

La *suma vectorial* de dos subespacios P y Q se define como el conjunto de todas las sumas de vectores pertenecientes a los subespacios P y Q . También se puede interpretar la suma vectorial como el subespacio engendrado por la unión de las bases de P y Q .

La *intersección* de dos subespacios se define como el conjunto de vectores que pertenecen a ambos subespacios simultáneamente. Por ejemplo, la suma vectorial de dos planos (es decir, de dos subespacios bidimensionales) en el espacio tridimensional ordinario es el espacio total (supuesto solamente que los dos planos no coincidan), y su intersección, una recta (hecha la misma salvedad).

Las dimensiones p y q de dos subespacios, la dimensión t de su suma vectorial, y la dimensión s de su intersección satisfacen la interesante relación siguiente:

$$p + q = t + s.$$

Omitimos la demostración de esta proposición.

De esta relación podemos hacer algunas deducciones referentes a la intersección de subespacios en casos especiales. Por ejemplo, dos planos no coincidentes (es decir, dos subespacios bidimensionales) en un espacio de cuatro dimensiones tienen en general una intersección que consta de un solo punto (la dimensión de su intersección es cero) y su intersección es una recta solamente si su suma vectorial es tridimensional, es decir, si ambos planos pertenecen al mismo subespacio tridimensional. Basta hacer uso de la relación anterior, escribiendo $t + s = 2 + 2 = 4$, de donde se deduce que $s = 1$ sólo si $t = 3$.

Espacios lineales complejos

En la descripción del espacio de filas y del espacio lineal general, no hemos especificado la clase de números de que hablábamos al definir la multiplicación de un vector por un número. Como hemos partido de una generalización de los vectores ordinarios, es decir, de segmentos dirigidos en el espacio geométrico tridimensional, hemos estado considerando implícitamente números reales. Los espacios así contruidos, llamados espacios lineales reales, generalizan de la forma más natural el espacio tridimensional de vectores ordinarios; sin embargo, en las matemáticas actuales es preciso considerar a veces *espacios lineales complejos*. Entendemos por esto una colección de objetos para los cuales están definidas las operaciones de adición y multiplicación por un número complejo, operaciones que satisfacen los axiomas 1-8. Podemos tomar como ejemplo de espacio complejo el espacio de vectores fila cuyos elementos son números complejos arbitrarios.

Formalmente, la teoría de espacios complejos no difiere esencialmente de la teoría de espacios reales.

Sin embargo, un espacio complejo bidimensional no tiene una interpretación geométrica intuitiva. El hecho es que un espacio complejo se puede considerar también como real, ya que la multiplicación por un número real está definida por estarlo para números complejos, y los reales son parte de los complejos. Pero la dimensión de un espacio complejo considerado como real es justamente el doble de la que tiene como espacio complejo. Si $e_1, e_2, \dots, e_n$ es una base de un espacio complejo, podemos tomar como base del mismo espacio considerado como real, por ejemplo, los vectores

$$e_1, ie_1, e_2, ie_2, \dots, e_n, ie_n, \quad \text{donde} \quad i = \sqrt{-1}.$$

Por tanto, un espacio complejo bidimensional se puede considerar como un espacio real pero de dimensión cuatro.

Además, la teoría de espacios lineales no sufre cambios formales si sustituimos la colección de números por los que se multiplican los vectores por otro conjunto de números distinto del de los reales o los complejos, con la única condición siguiente: las operaciones aritméticas básicas (adición, sustracción, multiplicación y división) deben estar definidas, y los resultados de dichas operaciones con números del conjunto deben seguir perteneciendo al conjunto. Un conjunto de números que satisface esta condición se llama cuerpo numérico.

(Este concepto será estudiado con más detalle en el capítulo 19.) Podemos tomar como ejemplo de cuerpo numérico el cuerpo de los números racionales.

En algunas partes del álgebra próximas a la teoría de números se puede aplicar la teoría de espacios lineales sobre un cuerpo arbitrario, con buenos resultados.

El espacio euclidiano n -dimensional

Algunos conceptos importantes del espacio vectorial ordinario no han sido generalizados en la exposición precedente; en particular nos referimos a los conceptos de longitud de un vector y de ángulo entre vectores. Es bien sabido que en geometría analítica los problemas relacionados con intersecciones de rectas y planos, paralelismo, así como otros muchos, no hacen uso de estos conceptos. Las propiedades de un espacio cuya descripción no requiere los conceptos de longitud y ángulo se caracterizan porque son precisamente las que permanecen inalterables en una transformación afín (ver capítulo 3, §11). Por esta razón, los espacios en los que no se definen los conceptos de longitud de un vector y de ángulo entre vectores reciben el nombre de *espacios afines*.

Sin embargo, muchos problemas matemáticos requieren generalizaciones de estos conceptos a los espacios n -dimensionales. Estas generalizaciones se efectúan mediante una analogía con la teoría de vectores en el plano o en el espacio.

Consideremos primeramente el espacio real de vectores fila. Definimos la longitud del vector $X = (x_1, x_2, \dots, x_n)$ el número

$$|X| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

Definición muy natural, pues para $n = 2$ y $n = 3$ la longitud de un vector se calcula precisamente por esta fórmula en función de sus coordenadas respecto a unos ejes cartesianos.

El concepto de ángulo entre vectores se introduce de forma también natural mediante las siguientes consideraciones. En el plano y en el espacio el ángulo entre los vectores X e Y es el ángulo en el vértice A del triángulo de lados $AB = |X|$, $AC = |Y|$ y $BC = |X - Y|$.

En un espacio n -dimensional parece natural hacer lo mismo para definir el ángulo entre vectores, es decir, proceder como si pudiéramos «extraer» un par de vectores del espacio n -dimensional y los «colocáramos» en un plano conservando su longitud y el ángulo que forman. No obstante, una definición así carece de rigor; la existencia

de un triángulo ABC de lados $|X|$, $|Y|$ y $|X - Y|$ debe ser demostrada.

Dejando de lado esta falta de rigor, introducimos una fórmula para el cálculo de ángulos entre vectores. Por una conocida fórmula de trigonometría tenemos

$$BC^2 = AB^2 + AC^2 - 2AB \cdot AC \cos \phi;$$

de aquí,

$$\begin{aligned} \cos \phi &= \frac{|X|^2 + |Y|^2 - |X - Y|^2}{2|X| \cdot |Y|} = \\ &= \frac{x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_n^2 - (x_1 - y_1)^2 - \cdots - (x_n - y_n)^2}{2|X| \cdot |Y|} = \\ &= \frac{x_1 y_1 + \cdots + x_n y_n}{|X| \cdot |Y|}. \end{aligned}$$

Si conservamos la expresión «producto escalar» —como en el espacio tridimensional— para el producto de las longitudes de dos vectores por el coseno del ángulo que forman, resulta que el producto escalar de vectores se calcula por la fórmula

$$X \cdot Y = x_1 y_1 + \cdots + x_n y_n,$$

que para $n = 2$ y $n = 3$ coincide con las fórmulas conocidas del producto escalar de vectores ordinarios.

En sentido estricto, la expresión $x_1 y_1 + \cdots + x_n y_n$ debería tomarse como definición de producto escalar (ya que existe una falta de rigor en la definición del producto escalar por medio de ángulos), definiendo entonces el ángulo entre dos vectores por la fórmula

$$\cos \phi = \frac{X \cdot Y}{|X| \cdot |Y|}. \quad (3)$$

Esto es lo que haremos.

Para justificar esta definición de ángulo debemos probar que el valor absoluto del segundo miembro de (3) no es mayor que 1; es decir, que $(X \cdot Y)^2 \leq |X|^2 \cdot |Y|^2$.

Desarrollando esta desigualdad resulta

$$(x_1 y_1 + \cdots + x_n y_n)^2 < (x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2).$$

Esta fórmula se conoce con el nombre de fórmula de Cauchy-Bunjakovski y se puede probar directamente mediante un cálculo muy tedioso. Nosotros lo demostraremos por el siguiente razonamiento indirecto.

Primeramente mencionamos las siguientes propiedades de producto escalar:

- 1'. $X \cdot X = |X|^2 > 0$ para $X \neq 0$.
- 2'. $X \cdot Y = Y \cdot X$.
- 3'. $(cX) \cdot Y = c(X \cdot Y)$.
- 4'. $(X_1 + X_2) \cdot Y = X_1 \cdot Y + X_2 \cdot Y$.

Estas propiedades se deducen inmediatamente de la expresión del producto escalar en función de las coordenadas.

Introducimos ahora el vector $Y + tX$, donde t es un número real arbitrario. Tenemos $|Y + tX|^2 \geq 0$, ya que el cuadrado de la longitud de un vector no puede ser negativo. Pero por las propiedades del producto escalar $|Y + tX|^2 = |Y|^2 + 2tX \cdot Y + t^2 |X|^2$. Se sabe además que un trinomio de segundo grado es no negativo para todos los valores de la variable real t si sus raíces son complejas o iguales, es decir, si su discriminante es negativo ó 0. Pero el discriminante del trinomio $|Y|^2 + 2tX \cdot Y + t^2 |X|^2$ es igual a $4(X \cdot Y)^2 - 4|X|^2 |Y|^2$, de modo que $(X \cdot Y)^2 - |X|^2 |Y|^2 \leq 0$, lo cual es equivalente a la desigualdad de Cauchy-Bunjakovski.

De esta desigualdad se deduce que

$$\frac{|X \cdot Y|}{|X| \cdot |Y|} \leq 1$$

y por tanto la definición de ángulo por la fórmula (3) está justificada.

Es fácil, además, deducir las desigualdades

$$|X| - |Y| \leq |X \pm Y| \leq |X| + |Y|,$$

que implican, en particular, la existencia de un triángulo de lados $|X|$, $|Y|$ y $|X - Y|$, de modo que la primera definición, no rigurosa, es ahora completamente válida.

Definición axiomática de un espacio euclídeo n-dimensional

En la sección precedente hemos introducido los conceptos de longitud de un vector, de ángulo y de producto escalar en el espacio de los vectores fila. En la definición axiomática general, estos conceptos se definen también axiomáticamente, comenzando por el producto escalar.

Multiplicación escalar de vectores es el nombre de la operación que asocia a cada par de vectores X e Y un número llamado producto escalar que se representa por $X \cdot Y$, y que cumple los siguientes postulados (axiomas):

- 1'. $X \cdot X > 0$ para $X \neq 0$, $0 \cdot 0 = 0$.
- 2'. $X \cdot Y = Y \cdot X$.
- 3'. $(cX) \cdot Y = c(X \cdot Y)$.
- 4'. $(X_1 + X_2) \cdot Y = X_1 \cdot Y + X_2 \cdot Y$.

Además, por longitud de un vector entendemos el número $\sqrt{X \cdot X}$, y por coseno del ángulo entre dos vectores, el cociente $X \cdot Y / |X| \cdot |Y|$. Para justificar esta última desigualdad, es necesario establecer la de Cauchy-Bunjakovski $(X \cdot Y)^2 \leq |X|^2 |Y|^2$. Pero se puede hacer exactamente igual que en la sección precedente. En nuestra demostración hemos usado solamente las propiedades 1', 2', 3', 4' del producto escalar; la naturaleza específica del espacio de filas no desempeñaba ningún papel. Un espacio lineal real con un producto escalar que satisfaga los axiomas 1' - 4' recibe el nombre de *espacio euclidiano*.

En los diversos espacios lineales que se estudian en matemáticas el producto escalar se introduce por diversos métodos, según la naturaleza del problema. Por ejemplo, en el espacio de las funciones de una variable $X(t)$ definidas en un intervalo $a \leq t \leq b$, se toma a menudo como producto escalar de dos funciones $X(t)$ e $Y(t)$ el número $\int_a^b X(t) Y(t) dt$ ó $\int_a^b X(t) Y(t) p(t) dt$, donde $p(t)$ es una función positiva. Es fácil comprobar que las dos definiciones satisfacen los axiomas 1' - 4'.

Ortogonalidad; bases ortogonales

Se dice que dos vectores de un espacio euclidiano son *ortogonales* (o perpendiculares) si su producto escalar es cero. Es fácil comprobar que un conjunto de vectores distintos del vector 0 y ortogonales dos a dos son linealmente independientes. Para ello supongamos que

Por las propiedades del producto escalar $X_1(c_1X_1 + c_2X_2 + \cdots + c_nX_n) = c_1 |X_1|^2 = 0$; de donde $c_1 = 0$. De la misma forma podemos probar que $c_2 = c_3 = \cdots = c_n = 0$. Por tanto $X_1, \cdots, X_n$ son linealmente independientes.

No es difícil probar, pero no lo haremos aquí, que un espacio euclidiano tiene un número infinito de bases ortogonales. Además, dado un subespacio P de un espacio R , podemos ampliar cualquier base de P añadiendo algunos vectores hasta formar una base de R .

$$X = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n \quad \text{e} \quad Y = y_1 e_1 + y_2 e_2 + \cdots + y_n e_n,$$
[illegible]

Así, la longitud de un vector y el producto escalar se expresan, en función de coordenadas en una base ortonormal, con la misma fórmula que en el espacio de filas.

La transición de uno de los modelos de espacio euclidiano, el espacio de filas, al espacio euclidiano general definido axiomática-

mente no introduce complicaciones, y además extiende el campo de aplicación de la teoría.

Veamos ahora el problema de la proyección ortogonal de un vector sobre un subespacio. Sea R_n un espacio euclidiano n -dimensional y P_m un subespacio de dimensión m . Sea además $e_1, e_2, \dots, e_m, f_1, \dots, f_{n-m}$ una base ortonormal de R_n que incluye una base de P_m . El subespacio Q_{n-m} generado por los vectores $f_1, f_2, \dots, f_{n-m}$ se llama *complemento ortogonal* del subespacio P_m . Su dimensión es $n - m$ y se caracteriza por ser el subespacio formado por todos los vectores ortogonales a cada vector de P_m .

Todo vector Z de R_n se puede expresar de forma única como suma de dos vectores X e Y pertenecientes uno a P_m y el otro a Q_{n-m} . Esto es claro, pues Z se puede expresar unívocamente de la forma

$$Z = x_1 e_1 + \dots + x_m e_m + y_1 f_1 + \dots + y_{n-m} f_{n-m},$$

siendo $X = x_1 e_1 + \dots + x_m e_m$, $Y = y_1 f_1 + \dots + y_{n-m} f_{n-m}$.

El vector X recibe el nombre de *proyección ortogonal* de Z sobre P_m .

Espacios unitarios

Los conceptos de longitud de un vector y de producto escalar de dos vectores se pueden definir también en un espacio complejo. En primer lugar, como en el caso anterior, se define el producto escalar de la forma siguiente: a cada par de vectores X e Y se le asocia un número complejo denotado por $X \cdot Y$ (no necesariamente real) y llamado producto escalar. La multiplicación escalar satisface los siguientes axiomas:

1''. $X \cdot X$ es real y positivo para $X \neq 0$, $0 \cdot 0 = 0$.

2''. $Y \cdot X = (X \cdot Y)'$. Aquí la prima significa el complejo conjugado.

3''. $(cX) \cdot Y = c(X \cdot Y)$ cualquiera que sea el número complejo c .

4''. $(X_1 + X_2) \cdot Y = X_1 \cdot Y + X_2 \cdot Y$ (ley distributiva).

En el espacio de filas con números complejos podemos tomar como producto escalar de los vectores $(x_1, x_2, \dots, x_n)$ e $(y_1, y_2, \dots, y_n)$ el número $x_1 y_1' + \dots + x_n y_n'$. Es fácil comprobar que esta definición satisface los axiomas 1'' - 4''.

Por definición, la longitud de un vector es el número $\sqrt{X \cdot X}$. No se define el concepto de ángulo entre dos vectores.

Un espacio lineal complejo con un producto escalar que satisface los axiomas 1'' - 4'' recibe el nombre de *espacio unitario*.

§3. *Sistemas de ecuaciones lineales**Sistemas de dos ecuaciones con dos incógnitas y de tres ecuaciones con tres incógnitas*

Un sistema de dos ecuaciones con dos incógnitas tiene la forma general siguiente:

$$\begin{aligned} a_1x + b_1y &= c_1, \\ a_2x + b_2y &= c_2. \end{aligned}$$

Para resolver este sistema multiplicamos la primera ecuación por b_2 , la segunda por $-b_1$ y sumamos. Obtenemos

$$(a_1b_2 - a_2b_1)x = c_1b_2 - c_2b_1.$$

Análogamente, multiplicando la primera ecuación por $-a_2$, la segunda por a_1 y sumando, obtenemos

$$(a_1b_2 - a_2b_1)y = a_1c_2 - a_2c_1.$$

A partir de estas ecuaciones es fácil determinar x e y siempre que la expresión $a_1b_2 - a_2b_1$ formada por los coeficientes de las incógnitas x e y sea diferente de cero. Esta expresión se llama *determinante* de la matriz

$$\begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix}$$

formada con los coeficientes del sistema. El determinante se escribe así:

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}$$

De la definición se deduce que el determinante se puede calcular por el esquema



que no requiere más explicaciones.

Volvamos a la solución del sistema. Las expresiones $c_1b_2 - c_2b_1$ y $a_1c_2 - a_2c_1$ también aparecen como determinantes de acuerdo con la definición, a saber,

$$\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix} \quad \text{y} \quad \begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}$$

Por tanto, si el determinante

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}$$

es distinto de cero, obtenemos las siguientes fórmulas para la solución del sistema:

$$x = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}; \quad y = \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}. \quad (4)$$

Estrictamente hablando, estos razonamientos no son completos. Las operaciones hechas con las ecuaciones para deducir la solución del sistema, tienen sentido solamente suponiendo que x e y son de hecho una solución del sistema. La base de nuestro razonamiento es la siguiente: si el determinante de los coeficientes del sistema es distinto de cero y la solución del sistema existe, entonces se puede calcular por las fórmulas (4). Es necesario, por tanto, verificar que los valores de las incógnitas hallados satisfacen de hecho las dos ecuaciones. Esto puede hacerse sin ninguna dificultad.

Así, si el determinante de la matriz de los coeficientes es distinto de 0, entonces el sistema tiene una solución única dada por las fórmulas (4).

Para un sistema de tres ecuaciones con tres incógnitas

$$\begin{aligned} a_1x + b_1y + c_1z &= d_1, \\ a_2x + b_2y + c_2z &= d_2, \\ a_3x + b_3y + c_3z &= d_3, \end{aligned}$$

la solución es fácil de hallar con razonamientos y cálculos similares; para esto es suficiente multiplicar las tres ecuaciones por factores convenientes de modo que luego al sumarlas desaparezcan dos incóg-

nit. Es fácil verificar mediante un sencillo cálculo que tomando los factores $b_2c_3 - b_3c_2$, $b_3c_1 - b_1c_3$ y $b_1c_2 - b_2c_1$ desaparecen las incógnitas y , z .

Como resultado obtenemos que, si la expresión

$$\Delta = a_1b_2c_3 - a_1b_3c_2 + a_2b_3c_1 - a_2b_1c_3 + a_3b_1c_2 - a_3b_2c_1$$

es diferente de cero, el sistema tiene una única solución que se obtiene mediante las fórmulas

$$x = \frac{\Delta_1}{\Delta}, \quad y = \frac{\Delta_2}{\Delta}, \quad z = \frac{\Delta_3}{\Delta},$$

donde Δ_1 , Δ_2 , Δ_3 se obtienen a partir de Δ , sustituyendo los coeficientes de la incógnita correspondiente por los términos independientes.

La expresión Δ se llama determinante de la matriz

$$\begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix}$$

y se denota por

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}$$

Los siguientes esquemas se emplean en el cálculo de determinantes:



En el primer esquema, las líneas (una diagonal y dos triángulos) relacionan la posición de los elementos cuyo producto entra en la expresión del determinante con signo positivo; en el segundo, las rectas relacionan los de signo negativo.

Para los dos tipos de sistemas estudiados, los resultados son completamente similares.

En ambos casos el sistema tiene solución única, supuesto que el

determinante de los coeficientes del sistema sea distinto de cero. Las fórmulas que dan la solución son también similares: en el denominador de cada incógnita está la matriz de los coeficientes, y en el numerador aparecen los determinantes que resultan de sustituir, en la matriz de coeficientes, los correspondientes a la incógnita que se va a calcular por los términos independientes.

La generalización inmediata de estos resultados a sistemas de n ecuaciones con n incógnitas tiene sus dificultades. En cambio, se hace relativamente fácil por un método indirecto: generalizamos primero el concepto de determinante a matrices cuadradas de cualquier orden y, una vez estudiadas sus propiedades, se aplican los resultados al estudio de sistemas de ecuaciones.

Determinantes de orden n

Cuando consideramos la expresión explícita de los determinantes

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1 b_2 - a_2 b_1$$

y

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1 b_2 c_3 - a_1 b_3 c_2 + a_2 b_3 c_1 - a_2 b_1 c_3 + a_3 b_1 c_2 - a_3 b_2 c_1$$

salta a la vista que cada término del desarrollo tiene un elemento de cada fila y otro de cada columna, y que en el determinante aparecen todos los posibles productos de esta forma, afectados de signo más o menos. En esta propiedad se basa la generalización del concepto de determinante a matrices cuadradas de orden arbitrario. El determinante de una matriz cuadrada de orden n , o más brevemente, un determinante de orden n , se define como la suma algebraica de todos los productos posibles de los elementos de la matriz, tomando un solo elemento de cada fila y un solo elemento de cada columna; estos productos van afectados de un signo más o menos según una regla bien definida. Esta regla es algo complicada de expresar y no nos detenemos en su formulación. Basta mencionar que garantiza el cumplimiento de las siguientes propiedades:

1. Al intercambiar dos filas, el determinante cambia de signo. Para determinantes de orden 2 y 3 es fácil de verificar mediante un cálculo inmediato. En el caso general se prueba haciendo uso de la regla de los signos, que no hemos formulado aquí.

Los determinantes tienen un gran número de propiedades que los hace útiles en muchos cálculos teóricos y numéricos a pesar de que su empleo es extraordinariamente engorroso: un determinante de orden n tiene, como es fácil de probar, $n!$ términos; cada término consta de n factores y está afectado de un signo de acuerdo con cierta regla.

Procedemos a enumerar las propiedades básicas de los determinantes, pero omitiendo sus demostraciones detalladas. La primera de ellas ha sido ya formulada.

2. Un determinante no varía al trasponer su matriz, es decir, al intercambiar las filas y las columnas conservando el orden. La demostración se basa en el estudio detallado de la regla de los signos. Esta propiedad permite enunciar para las columnas todas las proposiciones referentes a filas.

3. Un determinante es función lineal de los elementos de cada fila (o columna). Con detalle,

$$\begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{i1} & \cdots & a_{in} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in}, \quad (5)$$

donde $A_{i1}, A_{i2}, \dots, A_{in}$ son expresiones que no dependen de los elementos de la fila i .

Esta propiedad se deduce evidentemente del hecho de que cada término tiene un factor de cada fila, en particular de la fila i .

La ecuación (5) se llama desarrollo del determinante respecto a los elementos de la fila i , y los coeficientes $A_{i1}, A_{i2}, \dots, A_{in}$ son los complementos algebraicos de los elementos $a_{i1}, a_{i2}, \dots, a_{in}$ del determinante.

4. El complemento algebraico A_{ij} del elemento a_{ij} es igual, salvo el signo, al llamado menor Δ_{ij} del determinante, es decir, el determinante de orden $(n-1)$, que resulta de suprimir la fila i y la columna j del determinante dado. Para que el menor sea igual al complemento algebraico ha de ser tomado con el signo $(-1)^{i+j}$. Las propiedades 3 y 4 reducen el cálculo de un determinante de orden n al de n determinantes de orden $(n-1)$.

Las propiedades enumeradas tienen consecuencias interesantes. Mencionemos algunas de ellas.

5. Un determinante con dos filas iguales vale cero.

Basta observar que si tiene dos filas iguales no varía al intercambiar dichas filas, ya que son iguales; pero por otra parte debe cambiar de signo por la primera propiedad. Por tanto es igual a cero.

6. La suma de los productos de los elementos de una fila por los complementos algebraicos de otra es cero.

Una suma tal es el resultado de desarrollar un determinante con dos filas iguales respecto a una de ellas.

7. Si se multiplican los elementos de una fila por un número, el determinante queda multiplicado por dicho número.

Esto se sigue de la propiedad 3.

8. Un determinante con dos filas proporcionales es igual a cero.

Basta sacar fuera del determinante el factor de proporcionalidad; queda así otro con dos filas iguales.

9. Un determinante no cambia si sumamos a los elementos de una fila números proporcionales a los elementos de otra fila.

Por la propiedad 3, el determinante modificado es la suma del determinante original y de otro con dos filas proporcionales, que es cero.

La última propiedad proporciona un buen método para el cálculo de determinantes. Sin cambiar su valor, podemos transformar la matriz aplicando esta regla, de forma que todos los elementos de una fila (o columna), excepto uno, sean cero. Desarrollando entonces el determinante por los elementos de esa fila (o columna) reducimos su cálculo al de un único determinante de orden $n - 1$, a saber, el complemento algebraico del único elemento distinto de cero de la fila en cuestión.

Por ejemplo, supongamos que queremos calcular el determinante

$$\Delta = \begin{vmatrix} 1 & 1 & -1 & 2 \\ 2 & -1 & 1 & 1 \\ -1 & 2 & 0 & 1 \\ 1 & 1 & -2 & 1 \end{vmatrix}.$$

Sumamos a la segunda columna la primera multiplicada por -1 ; a la tercera, la primera, y a la cuarta, la primera multiplicada por -2 , y obtenemos

$$A = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 2 & -3 & 3 & -3 \\ -1 & 3 & -1 & 3 \\ 1 & 0 & -1 & -1 \end{vmatrix}.$$

Desarrollando Δ por los elementos de la primera fila, obtenemos

$$\Delta = 1 \cdot (-1)^{1+1} \begin{vmatrix} -3 & 3 & -3 \\ 3 & -1 & 3 \\ 0 & -1 & -1 \end{vmatrix}.$$

Finalmente, sumamos a la primera fila la segunda y desarrollamos respecto a los elementos de la primera columna, y obtenemos

$$\Delta = \begin{vmatrix} 0 & 2 & 0 \\ 3 & -1 & 3 \\ 0 & -1 & -1 \end{vmatrix} = 3 \cdot (-1)^{1+2} \begin{vmatrix} 2 & 0 \\ -1 & -1 \end{vmatrix} = -3 \cdot (-2) = 6.$$

El determinante de la matriz A se denota por $|A|$.

Como conclusión, mencionemos una propiedad muy importante de los determinantes.

El determinante del producto de dos matrices cuadradas es igual al producto de los determinantes de los factores, es decir, en notación abreviada, $|AB| = |A| |B|$.

Esta propiedad permite multiplicar determinantes de igual orden por la regla de multiplicación de matrices.

Sistemas de n ecuaciones lineales con n incógnitas

Con los métodos de la teoría de determinantes resulta muy fácil generalizar los resultados obtenidos para sistemas de dos y tres ecuaciones con dos y tres incógnitas, respectivamente, a sistemas de n ecuaciones con n incógnitas, suponiendo que el determinante de la matriz de coeficientes sea distinto de cero.

Sea

$$\begin{array}{l} a_{11}x_1 + \cdots + a_{1j}x_j + \cdots + a_{1n}x_n = b_1, \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nj}x_j + \cdots + a_{nn}x_n = b_n \end{array}$$

un sistema tal. Denotamos por Δ el determinante de la matriz de coeficientes del sistema, que suponemos distinto de cero. Denotamos además por A_{ij} el complemento algebraico de a_{ij} . Multiplicamos la primera ecuación por A_{1j} , la segunda por A_{2j} , $\dots$, la n -ésima por A_{nj} , y sumamos. Obtenemos

$$\Delta x_j = b_1 A_{1j} + \dots + b_n A_{nj}.$$

Basta observar que los coeficientes de todas las incógnitas excepto la x_j valen cero, pues son la suma de los productos de los complementos algebraicos de la columna j por los elementos de otras columnas (propiedad 6, aplicada a columnas); pero el coeficiente de la incógnita x_j es igual a la suma de los productos de los elementos de la columna j por sus complementos algebraicos, es decir, es igual a Δ .

Por tanto,

$$x_j = \frac{b_1 A_{1j} + \dots + b_n A_{nj}}{\Delta} \quad \text{para todo } j = 1, 2, \dots, n. \quad (6)$$

Como ya hemos dicho, estos razonamientos son válidos sólo si sobreentendemos que $x_1, x_2, \dots, x_n$ es una solución del sistema, debiendo suponer de antemano su existencia.

El resultado de la argumentación es el siguiente: si existe una solución, se puede hallar por las fórmulas (6) y es además única.

Para que la exposición sea completa, es necesario probar la existencia de una solución; esto se puede hacer sustituyendo los valores encontrados para las incógnitas en todas las ecuaciones de sistema original. Es fácil demostrar, haciendo uso de la misma propiedad de los determinantes (esta vez para filas), que efectivamente satisfacen todas las ecuaciones.

Así, el siguiente teorema es verdadero: si el determinante de la matriz de los coeficientes de un sistema de n ecuaciones con n incógnitas es distinto de cero, el sistema tiene una única solución, dada por las fórmulas (6).

Estas fórmulas se pueden transformar teniendo en cuenta que la suma $b_1 A_{1j} + \dots + b_n A_{nj}$ se puede escribir en forma de determinante:

$$\Delta_j = b_1 A_{1j} + \dots + b_n A_{nj} = \begin{vmatrix} a_{11} & \dots & b_1 & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & b_n & \dots & a_{nn} \end{vmatrix}$$

(los términos independientes están en la columna j).

Por tanto, hemos generalizado los resultados obtenidos para sistemas de dos y tres incógnitas a sistemas de n ecuaciones con n incógnitas; las fórmulas que dan la solución son formalmente las mismas.

He aquí un corolario de este teorema: si un sistema de ecuaciones no tiene solución o su solución no es única, el determinante de la matriz de los coeficientes es igual a cero.

Este corolario es de uso particularmente importante en los sistemas homogéneos, es decir, aquellos en los que los términos independientes $b_1, b_2, \dots, b_n$ son todos cero. Los sistemas homogéneos tienen siempre la solución «trivial» $x_1 = x_2 = \dots = x_n = 0$.

Si un sistema homogéneo tiene, además de la solución trivial, otra solución, su determinante es igual a cero.

Este último enunciado abre la posibilidad de usar la teoría de determinantes en otras ramas de las matemáticas y sus aplicaciones.

Consideremos, por ejemplo, un problema de geometría analítica: hallar la ecuación del plano que pasa por tres puntos dados (x_1, y_1, z_1) , (x_2, y_2, z_2) , (x_3, y_3, z_3) no alineados.

Por geometría elemental, se sabe que tal plano existe. Supongamos que su ecuación es de la forma $Ax + By + Cz + D = 0$. Entonces

$$\begin{aligned} Ax_1 + By_1 + Cz_1 + D &= 0, \\ Ax_2 + By_2 + Cz_2 + D &= 0, \\ Ax_3 + By_3 + Cz_3 + D &= 0. \end{aligned}$$

Sean x, y, z las coordenadas de un punto arbitrario de ese plano. Entonces también tenemos

$$Ax + By + Cz + D = 0.$$

Consideremos las cuatro ecuaciones como un sistema lineal homogéneo en las cuatro incógnitas A, B, C, D , coeficientes del plano buscado. Este sistema tiene una solución no trivial, pues el plano existe. Por tanto, el determinante del sistema es cero; es decir,

$$\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x & y & z & 1 \end{vmatrix} = 0 \quad (7)$$

Esta es, precisamente, la ecuación del plano, ya que es una ecuación de primer grado en x, y, z , hecho que se sigue de la linealidad del determinante respecto a los elementos de la última fila.

Partiendo de que los puntos dados no están alineados, es fácil verificar que no todos los coeficientes de esta ecuación son cero. En consecuencia, la ecuación (7) es efectivamente la ecuación de un plano. Y este plano pasa por los puntos dados, ya que sus coordenadas satisfacen evidentemente la ecuación.

Notación matricial para un sistema de n ecuaciones con n incógnitas

Un sistema de n ecuaciones con n incógnitas

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1, \\ &\dots\dots\dots \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= b_n. \end{aligned}$$

se puede escribir en notación matricial mediante una sola ecuación

$$AX = B.$$

Aquí, A denota la matriz de los coeficientes; X , la columna formada por las incógnitas, y B , la columna de los términos independientes.

La solución del sistema (si el determinante de la matriz A es distinto de cero) se puede escribir explícitamente como sigue [ver fórmula (6)]:

$$\begin{aligned} x_1 &= \frac{A_{11}}{\Delta} b_1 + \frac{A_{21}}{\Delta} b_2 + \cdots + \frac{A_{n1}}{\Delta} b_n, \\ x_2 &= \frac{A_{12}}{\Delta} b_1 + \frac{A_{22}}{\Delta} b_2 + \cdots + \frac{A_{n2}}{\Delta} b_n, \\ &\dots\dots\dots \\ x_n &= \frac{A_{1n}}{\Delta} b_1 + \frac{A_{2n}}{\Delta} b_2 + \cdots + \frac{A_{nn}}{\Delta} b_n \end{aligned}$$

o en forma de matriz

$$X = \begin{bmatrix} \frac{A_{11}}{\Delta} & \frac{A_{21}}{\Delta} & \cdots & \frac{A_{n1}}{\Delta} \\ \frac{A_{12}}{\Delta} & \frac{A_{22}}{\Delta} & \cdots & \frac{A_{n2}}{\Delta} \\ \dots\dots\dots \\ \frac{A_{1n}}{\Delta} & \frac{A_{2n}}{\Delta} & \cdots & \frac{A_{nn}}{\Delta} \end{bmatrix} B.$$

El primer factor del segundo miembro es una matriz llamada *inversa* de la matriz A y se denota por A^{-1} . Haciendo uso de esta notación obtenemos la solución del sistema de una forma sencilla y natural que recuerda a la solución de una ecuación con una sola incógnita:

$$X = A^{-1}B.$$

Podemos dar fácilmente otra demostración del resultado obtenido, recurriendo al álgebra de matrices.

Para ello debemos mencionar de entrada el papel especial que juega la matriz

$$E = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

llamada *matriz unidad*.

La matriz unidad tiene el mismo papel entre las matrices que el número 1 entre los números. Para cada matriz A se cumplen las siguientes ecuaciones: $AE = A$ y $EA = A$. Esto es fácil de comprobar mediante la regla de multiplicación de matrices.

La matriz A^{-1} definida anteriormente, inversa de A , tiene un significado similar al inverso de un número:

$$AA^{-1} = A^{-1}A = E.$$

La validez de estas ecuaciones se comprueba por la regla de multiplicación de matrices y por las propiedades 3 y 6 de los determinantes.

Conocidas las propiedades de las matrices unidad e inversa, se puede obtener la solución del sistema $AX = B$ por el siguiente método.

Supongamos que $AX = B$. Entonces $A^{-1}(AX) = A^{-1}B$. Pero $A^{-1}(AX) = (A^{-1}A)X = EX = X$ y por tanto $X = A^{-1}B$.

Supongamos ahora que $X = A^{-1}B$, entonces $AX = AA^{-1}B = EB = B$.

Así, la «ecuación» $AX = B$ tiene la solución única $X = A^{-1}B$, supuesto la existencia de A^{-1} . Para ello la única condición es que el determinante de la matriz A sea distinto de cero. Esta condición es no sólo suficiente sino también necesaria. Para probarlo supongamos que la matriz A tiene una inversa A^{-1} tal que $AA^{-1} = E$. Entonces,

por la propiedad del determinante del producto de matrices, tenemos que

$$|A| |A^{-1}| = |E| = 1,$$

y de aquí se sigue que el determinante de A no es cero.

Una matriz cuyo determinante no es cero se llama *no degenerada* o *no singular*. Hemos establecido, por tanto, que la inversa de una matriz existe si, y sólo si, es no degenerada.

La introducción del concepto de inversa de una matriz está justificada por su gran utilidad no sólo en la teoría de sistemas de ecuaciones lineales sino también en otros muchos problemas de álgebra lineal.

Como conclusión queremos advertir que las fórmulas encontradas para la resolución de sistemas lineales son indispensables en las consideraciones teóricas, pero no son convenientes para las soluciones numéricas. Como ya hemos mencionado, existen diversos métodos y esquemas de cálculo para la solución de sistemas numéricos; y en vista de la gran importancia de este problema para los trabajos prácticos, en la actualidad se sigue investigando sobre la simplificación de las soluciones numéricas de sistemas (especialmente con un gran número de incógnitas).

El caso general de sistemas de ecuaciones lineales

Vamos a estudiar ahora los sistemas de ecuaciones en el caso más general, es decir, cuando no se supone de antemano que el número de ecuaciones es igual al número de incógnitas. En una situación tan general no se puede esperar, naturalmente, que exista siempre una solución, o que si existe sea única. Parece lógico suponer que si el número de ecuaciones es menor que el de incógnitas, el sistema tenga infinitas soluciones. Por ejemplo, dos ecuaciones con tres incógnitas se satisfacen por las coordenadas de los puntos de la recta intersección de los planos definidos por las ecuaciones. Puede ocurrir, sin embargo, que el sistema no tenga solución, a saber, cuando los planos son paralelos. Y si el número de ecuaciones es mayor que el de incógnitas, cabe esperar que el sistema no tenga solución. No obstante, en este caso puede suceder que el sistema tenga soluciones incluso en número infinito.

Para estudiar la existencia y carácter de las soluciones de un sistema en el caso general, haremos uso de la interpretación «geométrica» del sistema.

jos, podemos dar una interpretación parecida en un espacio unitario, y obtener resultados similares.)

Sean $A'_1, A'_2, \dots, A'_m, Y$ vectores de un espacio euclidiano cuyas coordenadas, son respectivamente,

$$(a_{11}, a_{12}, \dots, a_{1n}), (a_{21}, a_{22}, \dots, a_{2n}), \dots, (a_{m1}, a_{m2}, \dots, a_{mn}) \\ (y_1, y_2, \dots, y_n)$$

Entonces el sistema toma la forma

$$A'_1 Y = 0, A'_2 Y = 0, \dots, A'_m Y = 0;$$

es decir, toda solución del sistema determina un vector ortogonal a todos los vectores formados por los coeficientes de las distintas ecuaciones.

Por tanto, el conjunto de soluciones forma un subespacio que es complemento ortogonal del subespacio generado por $A'_1, A'_2, \dots, A'_n$. La dimensión de este último subespacio es igual al rango r de la matriz formada con los coeficientes del sistema. La dimensión del complemento ortogonal, es decir, del «espacio de soluciones», es igual a $n - r$.

Ahora bien, todo subespacio tiene una base, es decir, un sistema de vectores linealmente independientes igual, en número, a la dimensión, y tal que sus combinaciones lineales generan todo el subespacio. Por tanto, entre las soluciones de un sistema homogéneo existen $n - r$ soluciones linealmente independientes, de forma que el resto de las soluciones son combinaciones lineales de ellas. Aquí n denota el número de incógnitas, y r , el rango de la matriz de coeficientes.

Así, pues, la estructura de las soluciones de un sistema homogéneo, y en consecuencia, también de un sistema no homogéneo, está completamente esclarecida. En particular, un sistema homogéneo tiene solamente la solución trivial $x_1 = x_2 = \dots = x_n = 0$ si, y sólo si, el rango de la matriz de los coeficientes es igual al número de incógnitas. Por lo que hemos dicho anteriormente, esta condición es también la condición de unicidad para la solución de un sistema no homogéneo (supuesta la condición de existencia).

El estudio de estos sistemas muestra cómo la introducción de conceptos geométricos generalizados proporciona sencillez y claridad a un complicado problema de álgebra.

$$\begin{aligned} y_1 &= \phi_1(x_1, x_2, \dots, x_n), \\ y_2 &= \phi_2(x_1, x_2, \dots, x_n), \\ &\vdots \\ y_n &= \phi_n(x_1, x_2, \dots, x_n), \end{aligned}$$
$$\begin{array}{ccc} \frac{\partial \phi_1}{\partial x_1} & \dots & \frac{\partial \phi_1}{\partial x_n} \\ \frac{\partial \phi_2}{\partial x_1} & \dots & \frac{\partial \phi_2}{\partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial \phi_n}{\partial x_1} & \dots & \frac{\partial \phi_n}{\partial x_n} \end{array}$$
$$\begin{aligned} y_1 &= a_{11}x_1 + \cdots + a_{1n}x_n, \\ y_2 &= a_{21}x_1 + \cdots + a_{2n}x_n, \\ &\vdots \\ y_n &= a_{n1}x_1 + \cdots + a_{nn}x_n, \end{aligned} \quad (11)$$

Una transformación lineal se puede definir también por sus propiedades intrínsecas, independientemente de las fórmulas (11) que

relacionan las coordenadas de puntos correspondientes. Una definición tal, en la que no se hace referencia a las coordenadas, es muy conveniente porque no depende de la elección de una base. Su definición es como sigue.

Una transformación lineal de un espacio lineal n -dimensional es una función $Y = A(X)$, donde X e Y son vectores. Esta función satisface el postulado de linealidad

$$A(c_1X_1 + c_2X_2) = c_1A(X_1) + c_2A(X_2) \quad (12)$$

En lo que sigue, cuando hablemos de transformaciones lineales lo haremos siempre en el sentido de esta definición.

Esta definición es equivalente a la que dimos antes en términos de las coordenadas. Pues sea la función $Y = A(X)$ que asocia al vector X , de coordenadas $x_1, x_2, \dots, x_n$, el vector Y , de coordenadas $y_1, y_2, \dots, y_n$, de forma que estas últimas se expresan en función de las primeras como en (11); es obvio que dicha función satisface el postulado (12). Inversamente, si la función $Y = A(X)$ satisface el postulado (12) y si $e_1, e_2, \dots, e_n$ es una base arbitraria del espacio, entonces

$$A(x_1e_1 + x_2e_2 + \dots + x_ne_n) = x_1A(e_1) + x_2A(e_2) + \dots + x_nA(e_n).$$

Denotemos las coordenadas (en la misma base) del vector $A(e_j)$ por $a_{1j}, \dots, a_{nj}$; $j = 1, \dots, n$. Entonces las coordenadas del vector $Y = A(X)$ son

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n, \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n, \\ &\dots\dots\dots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n. \end{aligned}$$

Así, a cada transformación lineal de un espacio lineal corresponde una cierta matriz cuadrada respecto a una base dada. Esta transformación se puede escribir en lenguaje matricial en la forma $Y = AX$. Aquí X es la columna de coordenadas del vector original; Y , la columna de coordenadas del vector transformado, y A , la matriz de los coeficientes de la transformación. Las columnas de la matriz A están formadas por las coordenadas de los transformados de los vectores de la base. De acuerdo con la notación matricial, escribiremos a menudo la propia transformación lineal en la forma $Y = AX$, omitiendo los paréntesis.

De la fórmula

$$A(x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n) = x_1A(\mathbf{e}_1) + x_2A(\mathbf{e}_2) + \cdots + x_nA(\mathbf{e}_n)$$

se deduce que una transformación lineal aplica el espacio completo en el subespacio generado por los vectores $A(\mathbf{e}_1), \dots, A(\mathbf{e}_n)$. La dimensión de este subespacio es igual al rango del sistema de vectores $A(\mathbf{e}_1), A(\mathbf{e}_2), \dots, A(\mathbf{e}_n)$ o, lo que es lo mismo, al rango de la matriz formada por sus coordenadas, es decir, al rango de la matriz asociada a la transformación. Este subespacio coincide con el espacio total si, y sólo si, el rango de la matriz A es igual a n , es decir, si el determinante de A es distinto de cero. En este caso la transformación se llama *no singular* o *no degenerada*.

Sabemos por la teoría de sistemas de ecuaciones lineales que únicamente son inversibles las transformaciones no degeneradas y que las coordenadas del vector original se expresan en función de las del transformado mediante la fórmula $\mathbf{X} = A^{-1}\mathbf{Y}$.

Una transformación cuya matriz tiene determinante cero se llama *singular* o *degenerada*. Una transformación degenerada no es inversible, como se deduce de la teoría de transformaciones lineales, o más intuitivamente, porque transforma el espacio total en parte de él.

Como primer ejemplo de transformación no degenerada podemos tomar la transformación identidad, que aplica cada vector en sí mismo y tiene como matriz en cualquier base la matriz unidad E . Una transformación lineal es también la que viene dada por una semejanza que consiste en multiplicar todos los vectores del espacio por un mismo número. La matriz correspondiente tampoco depende de la elección de la base y es de la forma aE , donde a es el factor de semejanza.

Un caso importante de transformaciones no degeneradas son las transformaciones ortogonales. El concepto de *transformación ortogonal* tiene sentido cuando se aplica a espacios euclidianos y se define como una transformación lineal que conserva la longitud de los vectores. Es una generalización al espacio n -dimensional de la rotación alrededor del origen de coordenadas o de una rotación seguida de una simetría respecto a un plano que pasa por el origen.

Es fácil ver que las transformaciones ortogonales conservan no sólo la longitud, sino también el producto escalar de vectores, y que en consecuencia transforman una base ortonormal del espacio en un sistema de vectores unitarios ortogonales dos a dos, es decir, en otra base ortonormal.

La matriz asociada a una transformación ortogonal respecto a una base ortonormal tiene las siguientes propiedades específicas.

Primero, la suma de los cuadrados de los elementos de cada columna es 1, ya que estas sumas son los cuadrados de las longitudes de los vectores imagen de los vectores de la base. Segundo, la suma de los productos de elementos correspondientes de distintas columnas es igual a cero, ya que estas sumas son el producto escalar de vectores transformados de la base.

En notación matricial estas dos propiedades se expresan mediante una única fórmula

$$\bar{P}P = E.$$

donde P es la matriz de la transformación ortogonal (respecto a una base ortonormal), y $\bar{P}$ es su traspuesta, es decir, la matriz cuyas filas son las columnas de P en el mismo orden.

Para probarlo basta ver que los elementos de la diagonal de $\bar{P}P$ son, por la regla de multiplicación de matrices, iguales a la suma de los cuadrados de los elementos de las correspondientes columnas de P , y que los demás elementos son la suma de los productos de los elementos correspondientes tomados de distintas columnas de P .

Como ejemplo de transformación degenerada podemos tomar la proyección ortogonal de todos los vectores de un espacio euclidiano sobre algún subespacio de él (ver §2); pues esta transformación aplica el espacio total sobre una parte de él.

Transformación de coordenadas

Consideramos ahora el problema de la transformación de coordenadas en un espacio n -dimensional, es decir, el problema de averiguar cómo cambian las coordenadas de un vector al pasar de una base a otra.

Sea $e_1, e_2, \dots, e_n$ la base original y $f_1, f_2, \dots, f_n$ otra base arbitraria del espacio. Suponemos además que

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}$$

es una matriz cuyas columnas son las coordenadas de los vectores de la nueva base $f_1, f_2, \dots, f_n$ respecto a la original. La matriz C es evidentemente no degenerada, ya que los vectores de la nueva base

Denotamos por $x_1, x_2, \dots, x_n$ las coordenadas de un cierto vector X respecto a la base original y por $x'_1, x'_2, \dots, x'_n$ las coordenadas del mismo vector respecto a la nueva. Entonces $X = x'_1 f_1 + x'_2 f_2 + \dots + x'_n f_n$ y, por tanto, las coordenadas del vector X respecto a la base original forman la columna

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} c_{11}x'_1 + c_{12}x'_2 + \cdots + c_{1n}x'_n \\ c_{21}x'_1 + c_{22}x'_2 + \cdots + c_{2n}x'_n \\ \vdots \\ c_{n1}x'_1 + c_{n2}x'_2 + \cdots + c_{nn}x'_n \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ c_{21} & \cdots & c_{2n} \\ \vdots & \vdots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix}.$$

Las fórmulas que expresan la relación entre las coordenadas de un vector respecto a las bases nueva y original coinciden formalmente con las fórmulas que ligan las coordenadas de vectores correspondientes en una transformación lineal no degenerada. Esta circunstancia permite interpretar una transformación lineal homogénea dada abstractamente mediante una matriz no degenerada, bien como una transformación de coordenadas, bien como una transformación del espacio. En cada caso concreto, la elección de una de las dos posibilidades está determinada por el contexto del problema considerado.

Supongamos una transformación lineal de matriz A respecto a una base $e_1, e_2, \dots, e_n$. La columna Y de las coordenadas del vector transformado está relacionada con la columna X del original por la fórmula

$$Y = AX.$$

Supongamos ahora que hacemos un cambio de coordenadas de matriz C ; X' e Y' denotan, respectivamente, las columnas de coordenadas, respecto a la nueva base, del vector original y de su transformado; de modo que $X = CX'$, $Y = CY'$, y por tanto

$$Y' = C^{-1}Y = C^{-1}AX = C^{-1}ACX'.$$

Así, pues, la matriz de nuestra transformación respecto a la nueva base es $C^{-1}AC$.

Dos matrices A y B ligadas por la relación $B = C^{-1}AC$, donde C es una matriz no singular, se dice que son *semejantes*. Las matrices correspondientes a la misma transformación lineal respecto a distintas bases son matrices semejantes.

*Autovectores y autovalores * de una transformación lineal*

Una clase muy importante de transformaciones son las que operan del siguiente modo.

Sean $e_1, e_2, \dots, e_n$ vectores linealmente independientes. Supongamos que, por efecto de una transformación, quedan multiplicados por los números $\lambda_1, \lambda_2, \dots, \lambda_n$. Si tomamos los vectores anteriores como base del espacio, se puede describir la transformación mediante la matriz

$$\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Las transformaciones de esta clase tienen una interpretación geométrica muy sencilla (por supuesto, solamente para espacios reales y $n = 2$ ó $n = 3$). Si todos los números son positivos, podemos interpretar la transformación como una dilatación (o una compresión) del espacio en la dirección de los vectores $e_1, e_2, \dots, e_n$ con coeficientes $\lambda_1, \lambda_2, \dots, \lambda_n$. Si alguno de los λ_i es negativo, la deformación del espacio va acompañada de un cambio de dirección por parte de alguno de los vectores $e_1, e_2, \dots, e_n$. Finalmente, si, por ejemplo, $\lambda_1 = 0$, entonces se produce una proyección del espacio paralelamente a e_1 sobre el subespacio engendrado por $e_2, \dots, e_n$, con la consiguiente deformación en estas direcciones.

La clase de transformaciones que hemos considerado es muy importante porque a pesar de su simplicidad son muy frecuentes. De hecho se puede probar que toda transformación lineal que satisfaga ciertas condiciones, no muy restrictivas, es de este tipo; es decir, podemos encontrar una base apropiada para la cual dicha transformación se puede describir mediante una matriz diagonal.

Las restricciones a imponer sobre la transformación resultan

* Vectores propios y valores propios son expresiones que también se utilizan. (N. del T.)

especialmente claras cuando se trata de transformaciones lineales en un espacio complejo. En lo que sigue supondremos que es así.

Introducimos la siguiente definición.

Un vector X , distinto del vector cero, que bajo una transformación lineal se transforma en un vector λX colineal recibe el nombre de *autovector de la transformación*. En otros términos, un vector X distinto de cero es un autovector de la transformación lineal A si, y sólo si, $AX = \lambda X$. El número λ se llama *autovalor* de la transformación A .

Es obvio que, si en alguna base la matriz de una transformación es diagonal, los vectores de dicha base son autovectores, y los elementos de la diagonal, autovalores. Y a la inversa: si existe una base compuesta de autovectores de la transformación A , entonces la matriz de la transformación A es diagonal respecto a esta base y sus elementos son los autovalores correspondientes a los vectores de la base.

Vamos a estudiar las propiedades de los autovectores y autovalores. Para ello escribimos la definición de autovector en notación coordenada. Sea A una matriz correspondiente a la transformación A en una cierta base y X la columna de coordenadas del vector X en la misma base. La ecuación $AX = \lambda X$ en notación coordenada se escribe $AX = \lambda X$, o bien

$$(A - \lambda E)X = 0.$$

Desarrollando esta ecuación obtenemos el sistema

$$\begin{array}{ccccccc} (a_{11} - \lambda)x_1 & + & a_{12}x_2 & + \cdots + & a_{1n}x_n & = & 0, \\ a_{21}x_1 & + & (a_{22} - \lambda)x_2 & + \cdots + & a_{2n}x_n & = & 0, \\ \dots\dots\dots & & & & & & \\ a_{n1}x_1 & + & a_{n2}x_2 & + \cdots + & (a_{nn} - \lambda)x_n & = & 0. \end{array}$$

Podemos considerar este sistema como un sistema lineal y homogéneo en las incógnitas $x_1, x_2, \dots, x_n$. Nos interesa el caso en que el sistema tenga una solución distinta de la trivial, ya que las coordenadas de un autovector no tienen por qué ser todas igual a cero. Sabemos que la condición necesaria y suficiente para la existencia de una solución no trivial de un sistema lineal homogéneo es que el rango de la matriz de los coeficientes sea menor que el número de

incógnitas, y esto es equivalente a que el determinante del sistema sea cero:

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0.$$

Por tanto, todos los autovalores de la transformación A son raíces del polinomio $|A - \lambda E|$ e, inversamente, cada raíz del polinomio es un autovalor de la transformación, ya que cada raíz corresponde por lo menos a un autovector. El polinomio $|A - \lambda E|$ se llama *polinomio característico* de la matriz A . La ecuación $|A - \lambda E| = 0$ se llama *ecuación característica* o *ecuación secular*, y sus raíces, *números característicos* de la matriz¹.

Por el teorema fundamental del álgebra (capítulo 4), todo polinomio tiene al menos una raíz; de aquí que toda transformación lineal tiene al menos un autovalor y por tanto un autovector. Pero, claro está, es posible que, incluso en el caso de que una transformación se pueda expresar mediante una matriz real, todos sus autovalores sean complejos. En consecuencia, el teorema de existencia de autovalores y autovectores de una transformación lineal arbitraria no es cierto en un espacio real. Basta considerar, por ejemplo, en el plano, un giro de ángulo distinto de 180° alrededor del origen de coordenadas: cambia la dirección de todos los vectores, no habiendo, por tanto, ni autovectores ni autovalores.

Las raíces de la ecuación característica de una matriz A son los autovalores de la transformación A ; por tanto, las matrices correspondientes a una misma transformación en distintas bases tienen polinomios característicos con las mismas raíces. Esto conduce a la siguiente afirmación: el polinomio característico de una transformación lineal depende solamente de ella y no de la elección de la base. Esto se puede comprobar mediante un elegante cálculo basado en las propiedades de las operaciones con matrices y determinantes.

Sabemos que si A es la matriz de una transformación en una cierta base, en otra base distinta es la matriz $C^{-1}AC$, donde C es una matriz

¹ El nombre de «ecuación secular» nació en la mecánica celeste en relación con el problema de las llamadas perturbaciones seculares en los movimientos de los planetas.

no singular. Podemos efectuar entonces las operaciones siguientes:

$$\begin{aligned} |C^{-1}AC - \lambda E| &= |C^{-1}AC - C^{-1}\lambda EC| = |C^{-1}(A - \lambda E)C| = \\ &= |C^{-1}| |C| |A - \lambda E| = |C^{-1}C| |A - \lambda E| = |A - \lambda E|. \end{aligned}$$

Así, pues, las matrices correspondientes a la misma transformación en distintas bases tienen el mismo polinomio característico, que se puede llamar, por tanto, polinomio de la transformación.

Supongamos ahora que todos los autovalores de la transformación A son distintos. Se prueba entonces que los autovectores, uno por cada autovalor, son linealmente independientes. Para ello, supongamos que algunos de ellos, pongamos $e_1, \dots, e_k$, son linealmente independientes, y que el resto, entre ellos e_{k+1} , dependen linealmente de los primeros. Entonces,

$$e_{k+1} = c_1 e_1 + c_2 e_2 + \dots + c_k e_k. \quad (13)$$

Si aplicamos la transformación lineal a ambos miembros de la ecuación (13), obtenemos

$$Ae_{k+1} = c_1 Ae_1 + c_2 Ae_2 + \dots + c_k Ae_k$$

de donde se sigue, por la definición de autovector, que

$$\lambda_{k+1} e_{k+1} = c_1 \lambda_1 e_1 + c_2 \lambda_2 e_2 + \dots + c_k \lambda_k e_k.$$

Si multiplicamos la ecuación (13) por λ_{k+1} y restamos de ella la ecuación que acabamos de obtener, queda:

$$c_1(\lambda_{k+1} - \lambda_1)e_1 + c_2(\lambda_{k+1} - \lambda_2)e_2 + \dots + c_k(\lambda_{k+1} - \lambda_k)e_k = 0,$$

y como $e_1, e_2, \dots, e_k$ son linealmente independientes, ocurre que

$$c_1(\lambda_{k+1} - \lambda_1) = c_2(\lambda_{k+1} - \lambda_2) = \dots = c_k(\lambda_{k+1} - \lambda_k) = 0.$$

Pero como hemos supuesto que los autovalores son todos distintos y hemos elegido un vector por cada autovalor, tenemos $\lambda_{k+1} - \lambda_1 \neq 0$, $\lambda_{k+1} - \lambda_2 \neq 0$, $\dots$, $\lambda_{k+1} - \lambda_k \neq 0$, y la ecuación (13) es imposible, ya que los coeficientes $c_1, c_2, \dots, c_k$ no pueden ser todos cero.

Ahora está claro que, si todos los autovalores de una transforma-

ción lineal son distintos, existe una base en la cual la matriz de la transformación es diagonal. Basta tomar como tal base un sistema de autovectores, uno por cada autovalor; como ya hemos demostrado, son linealmente independientes y su número es igual a la dimensión del espacio; es decir, forman efectivamente una base.

El teorema que acabamos de probar también se puede establecer por la teoría de matrices, como sigue. Si todos los autovalores son distintos, la matriz es semejante a una matriz diagonal, cuyos elementos diagonales sean esos autovalores.

La reducción de la matriz de una transformación lineal a su forma más simple es un problema considerablemente más complicado cuando las raíces del polinomio característico no son todas distintas. Nos limitaremos a exponer el resultado final.

Llamamos «caja canónica» de orden m a una matriz de la forma

$$I_{m, \lambda_i} = \begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{bmatrix}.$$

Todos los términos no indicados son iguales a cero.

Llamamos matriz canónica de Jordan a una matriz cuya diagonal principal está compuesta por «cajas canónicas», siendo cero el resto de los elementos:

$$\begin{bmatrix} \boxed{I_{m_1, \lambda_1}} & & & \\ & \boxed{I_{m_2, \lambda_2}} & & \\ & & \ddots & \\ & & & \boxed{I_{m_k, \lambda_k}} \end{bmatrix}.$$

Los números λ_i de las distintas «cajas» no son necesariamente distintos dos a dos. Toda matriz se puede reducir a una matriz canónica de Jordan semejante a ella. La demostración de este teorema es algo complicada. Conviene decir que juega un importante papel en muchas aplicaciones del álgebra a otros problemas de matemá-

Una matriz se puede reducir a forma diagonal si, y sólo si, los órdenes de todas las cajas son iguales a 1.

Definición y propiedades sencillas

Una forma cuadrática en n variables $x_1, x_2, \dots, x_n$ consta de términos de dos tipos: cuadrados de las variables y productos de dos variables, ambos con ciertos coeficientes. Una forma cuadrática se puede escribir en el siguiente esquema cuadrático:

Los pares de términos semejantes $a_{12}x_1x_2$ y $a_{21}x_2x_1$, etc., llevan el mismo coeficiente, igual a la mitad del coeficiente total del producto. Así, cada forma lleva asociada de manera única una matriz simétrica, formada por sus coeficientes.

Podemos representar una forma cuadrática de manera muy sencilla usando la notación matricial. Denotamos por X la columna de las variables $x_1, x_2, \dots, x_n$ y por $\bar{X}$ la fila $(x_1, x_2, \dots, x_n)$, es decir, la matriz traspuesta de X . Entonces

[illegible]

Las formas cuadráticas tienen interés en muchas ramas de las matemáticas y sus aplicaciones.

En la teoría de números y en cristalografía se utilizan formas cuadráticas en las cuales las variables $x_1, x_2, \dots, x_n$ toman solamente valores enteros. En geometría analítica aparecen en el estudio de curvas (o superficies) de segundo orden. En mecánica física, en la expresión de la energía cinética de un sistema en función de las componentes de las velocidades generalizadas, etc. Además, las formas cuadráticas son necesarias, en análisis, para el estudio de funciones de varias variables en relación con problemas donde es importante saber cómo se desvía una función respecto de su función lineal aproximada en el entorno de un punto. El estudio de los máximos y mínimos de una función es un ejemplo de este tipo.

Consideremos, por ejemplo, el problema de encontrar los máximos y mínimos de una función de dos variables $w = f(x, y)$ que tiene derivadas parciales continuas de tercer orden. Una condición necesaria para que la función w tenga un máximo o mínimo en el punto (x_0, y_0) es que las derivadas primeras se anulen en ese punto. Supongamos que esta condición se cumple. Damos a las variables pequeños incrementos h y k y consideramos el correspondiente incremento de la variable $\Delta w = f(x_0 + h, y_0 + k) - f(x_0, y_0)$. Por la fórmula de Taylor, este incremento es igual (con un error menor que infinitésimos de orden superior) a la forma cuadrática $\frac{1}{2}(rh^2 + 2shk + tk^2)$, donde r, s, t son los valores de las derivadas segundas $\partial^2 w / \partial x^2$, $\partial^2 w / \partial x \partial y$, $\partial^2 w / \partial y^2$, calculadas en el punto (x_0, y_0) . Si esta forma cuadrática es positiva para todos los valores de h y k (excepto para $h = k = 0$), la función w tiene un mínimo en (x_0, y_0) ; si es negativa, tiene un máximo. Finalmente, si la forma toma tanto valores positivos como negativos, entonces no tiene ni máximo ni mínimo. Las funciones con un número mayor de variables se estudian de forma similar.

El estudio de las formas cuadráticas consiste, esencialmente, en el estudio de la equivalencia de formas en relación a ciertos conjuntos de transformaciones lineales de las variables. Se dice que dos formas cuadráticas son *equivalentes* si se puede transformar la una en la otra mediante una transformación del conjunto dado. Muy relacionado con este problema está el de la *reducción* de una forma, es decir, su transformación en una forma lo más simple posible.

En distintos problemas relacionados con formas cuadráticas consideraremos distintos conjuntos de transformaciones de las variables.

En los problemas de análisis se admite cualquier tipo de transformación no singular de variables; en geometría analítica el interés

Si $a_{11} \neq 0$ no se requiere ningún cambio. Si $a_{11} = 0$, pero alguno de los coeficientes de la diagonal $a_{kk} \neq 0$, ponemos $x_1 = x'_k$, $x_k = x'_1$,

es evidentemente no singular. Tras la realización de este cambio, la forma queda reducida a

$$a_{11}x_1'^2 + f_1(x_2', \dots, x_n')$$

Continuando el proceso de forma similar, reducimos la forma a la forma «canónica» pedida:

$$a_1z_1^2 + a_2z_2^2 + \dots + a_nz_n^2.$$

Aquí $z_1, z_2, \dots, z_n$ son las nuevas variables introducidas en el último paso.

Ley de inercia de una forma cuadrática

En la reducción de una forma cuadrática a su forma canónica existe una considerable arbitrariedad en la elección del cambio de variables. Esta arbitrariedad procede, entre otras cosas, de que antes de comenzar a separar cuadrados podemos hacer un cambio de variables no singular arbitrario.

Sin embargo, a pesar de esta arbitrariedad, obtenemos como resultado de la reducción una forma cuadrática canónica casi única en el siguiente sentido: en las nuevas variables, el número de cuadrados que tienen coeficientes positivos, negativos o nulos es siempre el mismo, independientemente del método de la reducción y de los cambios efectuados. Este teorema se conoce con el nombre de *ley de inercia* de las formas cuadráticas. No lo demostraremos.

La ley de inercia de una forma cuadrática resuelve el problema de la equivalencia de formas cuadráticas reales en relación a todos los cambios de variable no singulares; pues dos formas son equivalentes si, y sólo si, al reducirlas a su forma canónica tienen el mismo número de cuadrados con coeficientes positivos, negativos y nulos.

Con vistas a las aplicaciones son de especial interés las formas cuadráticas que en su forma canónica tienen todos los coeficientes de los cuadrados *positivos*. Una forma de este tipo se llama *definida positiva*.

Las formas cuadráticas definidas positivas se caracterizan porque, para los valores reales —no todos iguales a cero— de las variables, son siempre positivas.

Transformación ortogonal de formas cuadráticas a forma canónica

Entre los posibles métodos de reducir una forma cuadrática a su forma canónica tienen especial interés las transformaciones ortogonales, es decir, aquellas que se obtienen mediante una transformación lineal de variables con una matriz ortogonal. Tales transformaciones son de interés, por ejemplo, en geometría analítica para reducir la ecuación de una curva o superficie de segundo orden a su forma canónica.

Para convencernos de la posibilidad de tal transformación es conveniente considerar las formas cuadráticas como funciones de vectores de un espacio euclidiano, y las variables $x_1, x_2, \dots, x_n$ como las coordenadas de un vector variable en una base ortonormal. Entonces, un cambio ortogonal de variables se puede interpretar como el paso de una base ortonormal a otra.

Asociamos a la forma cuadrática

$$f(x_1, x_2, \dots, x_n) = a_{11}x_1^2 + \dots + a_{1n}x_1x_n + \dots + a_{n1}x_nx_1 + \dots + a_{nn}x_n^2$$

la transformación lineal A , cuya matriz respecto a la base elegida es

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}.$$

Podemos considerar así nuestra forma cuadrática como el producto escalar $AX \cdot X$ (donde X es el vector de coordenadas $x_1, x_2, \dots, x_n$), y sus coeficientes como el producto escalar $Ae_i \cdot e_j$, donde $e_1, e_2, \dots, e_n$ es la base ortonormal elegida.

Es fácil ver que, a consecuencia de la simetría de A , se verifica para todo par de vectores X e Y la ecuación

$$AX \cdot Y = X \cdot AY.$$

Probaremos antes que nada que la transformación A tiene al menos un autovalor real (con su autovector correspondiente).

Para ello consideramos los valores de la forma $AX \cdot X$, donde X

varía sobre la esfera unidad, es decir, en el conjunto de los vectores unitarios. En estas condiciones la forma $AX \cdot X$ debe tener un máximo. Vamos a probar que este máximo es un autovalor de A y que el vector X_0 para el que se alcanza dicho máximo es el autovector correspondiente; es decir $AX_0 = \lambda_1 X_0$.

La demostración se hace por un procedimiento indirecto, a saber, estableciendo que el vector AX_0 es ortogonal a todos los vectores ortogonales a X_0 .

Observemos que la desigualdad $AZ \cdot Z \leq \lambda_1 |Z|^2$ es válida cualquiera que sea el vector Z . Esto es obvio, ya que $X = Z/|Z|$ es un vector unitario y λ_1 es el valor máximo de $AX \cdot X$ en la esfera unidad. Consideremos el vector $Z = X_0 + \epsilon Y$, donde ϵ es un número real, e Y , un vector arbitrario ortogonal a X_0 . Entonces

$$AZ \cdot Z = (AX_0 + \epsilon AY)(X_0 + \epsilon Y) = AX_0 \cdot X_0 + 2\epsilon AX_0 \cdot Y + \epsilon^2 AY \cdot Y = \lambda_1 + 2\epsilon AX_0 \cdot Y + \epsilon^2 AY \cdot Y.$$

Además,

$$|Z|^2 = (X_0 + \epsilon Y)(X_0 + \epsilon Y) = |X_0|^2 + \epsilon^2 |Y|^2 = 1 + \epsilon^2 |Y|^2,$$

ya que

$$X_0 \cdot Y = 0, \quad |X_0|^2 = 1.$$

Por tanto,

$$\lambda_1 + 2\epsilon AX_0 \cdot Y + \epsilon^2 AY \cdot Y \leq \lambda_1 + \epsilon^2 \lambda_1 |Y|^2,$$

y de aquí, dividiendo por ϵ^2 ,

$$\frac{2}{\epsilon} AX_0 \cdot Y < \lambda_1 |Y|^2 - AY \cdot Y \quad (14)$$

Esta última desigualdad debe ser satisfecha para cualquier número real de valor absoluto suficientemente pequeño

Pero esto sólo puede ocurrir a condición de que $AX_0 \cdot Y = 0$, ya que si $AX_0 \cdot Y > 0$, la desigualdad (14) es imposible para valores positivos de ϵ suficientemente pequeños, y si $AX_0 \cdot Y < 0$, es imposible para un ϵ negativo de valor absoluto suficientemente pequeño. Así, pues, $AX_0 \cdot Y = 0$, es decir, AX_0 es ortogonal a todos los vectores

ortogonales a X_0 . Por tanto, AX_0 y X_0 son colineales; es decir, $AX_0 = \lambda' X_0$, donde λ' es un número real. Es fácil verificar que

$$\lambda_1 = AX_0 \cdot X_0 = \lambda' X_0 \cdot X_0 = \lambda'.$$

Podemos probar ahora que toda forma cuadrática se puede reducir a su forma canónica mediante una transformación ortogonal.

Sea $e_1, e_2, \dots, e_n$ la base ortonormal original del espacio y $f_1, f_2, \dots, f_n$ una nueva base también ortonormal en la que f_1 es un autovector X_0 de la transformación A . Sean $x_1, x_2, \dots, x_n$ las coordenadas del vector X en la base original, y $x'_1, x'_2, \dots, x'_n$ las coordenadas del mismo vector en la nueva. Entonces

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = P \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix},$$

donde P es una matriz ortogonal.

Efectuemos el paso a las nuevas variables en la forma cuadrática $AX \cdot X$. Los nuevos coeficientes de la forma cuadrática son $a'_{ij} = Af_i \cdot f_j$. Por tanto

$$a'_{11} = Af_1 \cdot f_1 = \lambda_1 f_1 \cdot f_1 = \lambda_1,$$

$$a'_{1j} = a'_{j1} = Af_1 \cdot f_j = \lambda_1 f_1 \cdot f_j = 0 \quad \text{para } j \neq 1;$$

es decir, la forma es ahora

$$\lambda_1 x_1'^2 + \phi(x'_2, \dots, x'_n).$$

Así, mediante una transformación ortogonal hemos conseguido separar un cuadrado de una nueva variable.

Repitiendo el mismo razonamiento con la nueva forma $\phi(x'_2, \dots, x'_n)$, etc., llegamos a la conclusión de que por una cadena de transformaciones ortogonales se puede reducir una forma cuadrática a su forma canónica. Pero es evidente que una cadena de transformaciones ortogonales es otra transformación ortogonal. Esto concluye la demostración del teorema.

§6. *Funciones de matrices y algunas de sus aplicaciones**Funciones de matrices*

Las aplicaciones del álgebra lineal a otras ramas de las matemáticas son muy diversas y numerosas. No es una exageración decir que sus ideas y resultados se aplican a una gran parte de las matemáticas y física contemporáneas en una u otra forma y particularmente en forma de cálculo matricial.

Veamos brevemente uno de los métodos de aplicación del cálculo matricial a la teoría de ecuaciones diferenciales ordinarias, donde las funciones de matrices juegan un importante papel.

En primer lugar definimos las potencias de una matriz cuadrada A . Ponemos $A^0 = E$, $A^1 = A$, $A^2 = AA$, $A^3 = A^2A$, $A^4 = A^3A$, etc. Es fácil probar por la ley asociativa que $A^m A^n = A^{m+n}$ para cualesquiera números naturales m y n . Las operaciones de adición y multiplicación por un número han sido definidas ya para matrices. Esto permite definir de forma natural el significado de un polinomio (de una variable) de una matriz: si $\phi(x) = a_0 x^n + x^{n-1} + \dots + a_n$ entonces (por definición) escribimos $\phi(A) = a_0 A^n + a_1 A^{n-1} + \dots + a_n E$. Hemos definido así la función más simple de una matriz: el polinomio.

Mediante el proceso de paso al límite se puede generalizar el concepto de función de una matriz a una clase de funciones considerablemente más amplia. Sin tratar el problema en toda su generalidad, nos vamos a restringir a las funciones analíticas.

Primeramente introducimos el concepto de límite de una sucesión de matrices.

Se dice que una sucesión de matrices

$$A_1 = \begin{bmatrix} a_{11}^{(1)} & \dots & a_{1n}^{(1)} \\ \dots & \dots & \dots \\ a_{n1}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix}, \quad A_2 = \begin{bmatrix} a_{11}^{(2)} & \dots & a_{1n}^{(2)} \\ \dots & \dots & \dots \\ a_{n1}^{(2)} & \dots & a_{nn}^{(2)} \end{bmatrix}, \dots$$

converge a la matriz

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$$

(o que la matriz A es su límite) si $\lim_{k \rightarrow \infty} a_{ij}^{(k)} = a_{ij}$ para todo i, j . Definimos además la suma de la serie $A_1 + A_2 + \cdots + A_k + \cdots$ como el límite de las sumas parciales $\lim_{k \rightarrow \infty} (A_1 + A_2 + \cdots + A_k)$, si este límite existe.

Sea $f(z)$ una función analítica en un entorno del origen. Entonces $f(z)$ se puede desarrollar en serie de potencias

$$f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_k z^k + \cdots.$$

Para cualquier matriz cuadrada A parece natural poner

$$f(A) = a_0 E + a_1 A + a_2 A^2 + \cdots + a_k A^k + \cdots.$$

Sucede ahora que esta serie converge para todas las matrices A cuyos autovalores pertenecen al círculo de convergencia de la serie de potencias $a_0 + a_1 z + \cdots + a_k z^k + \cdots$.

De especial interés para las aplicaciones son las funciones elementales de matrices.

Por ejemplo, la serie geométrica $E + A + A^2 + \cdots + A^k + \cdots$ es convergente para aquellas matrices cuyos autovalores tienen valor absoluto menor que 1, y su suma es la matriz $(E - A)^{-1}$, de completo acuerdo con la fórmula

$$1 + x + \cdots + x^k + \cdots = \frac{1}{1 - x}.$$

La representación de la matriz $(E - A)^{-1}$ en forma de serie infinita proporciona un método efectivo para aproximar soluciones de sistemas de ecuaciones lineales cuya matriz de coeficientes difiera poco de la matriz unidad.

Pues si escribimos un sistema tal en la forma

$$(E - A)X = B,$$

obtenemos

$$X = (E - A)^{-1} B = B + AB + A^2 B + \cdots, \quad (15)$$

que es una fórmula muy interesante para la solución del sistema, a condición solamente de que la serie (15) converja suficientemente de prisa.

También es interesante la serie binómica

$$(E + A)^m = E + \frac{m}{1} A + \frac{m(m-1)}{2!} A^2 + \dots,$$

que se puede aplicar (supuesto que los autovalores de A sean en valor absoluto menores que 1) no sólo para exponentes naturales, sino también racionales y negativos.

Particularmente importante por sus aplicaciones es la función exponencial

$$e^A = E + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots$$

Esta serie converge para toda matriz A . Sus propiedades recuerdan a las de la función exponencial ordinaria. Por ejemplo, si A y B conmutan para la multiplicación, es decir, $AB = BA$, entonces $e^{A+B} = e^A \cdot e^B$. Sin embargo, la fórmula deja de ser válida si A y B no conmutan.

Aplicación a la teoría de sistemas de ecuaciones diferenciales lineales ordinarias

En la teoría de sistemas de ecuaciones diferenciales lineales ordinarias es conveniente considerar matrices cuyos elementos son funciones de una variable:

$$U(t) = \begin{bmatrix} a_{11}(t) & \cdots & a_{1n}(t) \\ \cdots & \cdots & \cdots \\ a_{m1}(t) & \cdots & a_{mn}(t) \end{bmatrix}.$$

Para tales matrices la derivada respecto a la variable t se define del siguiente modo:

$$\frac{dU(t)}{dt} = \begin{bmatrix} a'_{11}(t) & \cdots & a'_{1n}(t) \\ \cdots & \cdots & \cdots \\ a'_{m1}(t) & \cdots & a'_{mn}(t) \end{bmatrix}.$$

No es difícil comprobar que algunas fórmulas elementales de diferenciación también son válidas para matrices. Por ejemplo,

$$\begin{aligned}\frac{d(U + V)}{dt} &= \frac{dU}{dt} + \frac{dV}{dt}, \\ \frac{d(cU)}{dt} &= c \frac{dU}{dt}, \\ \frac{d(UV)}{dt} &= \frac{dU}{dt} V + U \frac{dV}{dt}.\end{aligned}$$

(La multiplicación se debe efectuar estrictamente en el orden dado por la fórmula.)

Un sistema homogéneo de ecuaciones diferenciales lineales ordinarias

$$\begin{aligned}\frac{dy_1}{dt} &= a_{11}(t)y_1 + a_{12}(t)y_2 + \cdots + a_{1n}(t)y_n \\ \frac{dy_2}{dt} &= a_{21}(t)y_1 + a_{22}(t)y_2 + \cdots + a_{2n}(t)y_n \\ &\dots\dots\dots \\ \frac{dy_n}{dt} &= a_{n1}(t)y_1 + a_{n2}(t)y_2 + \cdots + a_{nn}(t)y_n\end{aligned}$$

se puede escribir en notación matricial de la forma

$$\frac{dY}{dt} = A(t)Y,$$

donde

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad A(t) = \begin{bmatrix} a_{11}(t) & \cdots & a_{1n}(t) \\ \dots\dots\dots \\ a_{n1}(t) & \cdots & a_{nn}(t) \end{bmatrix},$$

es decir, de forma similar a una única ecuación diferencial lineal y homogénea.

Si los coeficientes del sistema son constantes, es decir, si la matriz A es constante, la solución también se parece formalmente a la de la ecuación $y' = ay$. En este caso es $Y = e^{At}C$, donde C es una columna de constantes arbitrarias.

Para el cálculo efectivo, la solución en esta forma resulta muy conveniente. El hecho es que para una función analítica arbitraria $f(z)$ tenemos

$$f(B^{-1}LB) = B^{-1}f(L)B.$$

Puesto que toda matriz se puede reducir a la forma canónica de Jordan (ver §4), el cálculo de una función de una matriz arbitraria se reduce al cálculo de una función de una matriz canónica, fácil de hallar. Por tanto, si $A = B^{-1}LB$, donde L es su matriz canónica, entonces

$$Y = e^{At}C = B^{-1}e^{Lt}BC = B^{-1}e^{Lt}C',$$

donde $C' = BC$ es una columna de constantes arbitrarias.

De esta fórmula se deriva fácilmente una expresión explícita para los componentes de la columna Y .

El matemático soviético I. A. Lappo-Danilevski ha desarrollado con éxito el aparato matemático de la teoría de matrices funcionales; él fue el primero en aplicarla a la investigación de sistemas de ecuaciones diferenciales. Sus resultados se cuentan entre las realizaciones más brillantes de las matemáticas en los últimos cincuenta años.

Capítulo 16

GEOMETRIAS NO EUCLIDIANAS

Desde que N. I. Lobachevski demostró la posibilidad de una geometría no euclidiana y propuso nuevas ideas sobre la relación de la geometría con la realidad material, el alcance de la geometría, sus métodos y sus aplicaciones se han visto enormemente ampliados. En nuestros días los matemáticos estudian diversos «espacios»: aparte de los espacios euclidianos tratan con espacios de Lobachevski, espacios proyectivos, distintos espacios n -dimensionales e incluso de infinitas dimensiones, espacios riemannianos, topológicos y muchos otros; el número de tales espacios es ilimitado y cada uno de ellos tiene su propia geometría y propiedades especiales. En física se utilizan los llamados «espacios» de las fases y de configuración; la teoría de la relatividad emplea espacios curvados y otros resultados de teorías geométricas abstractas.

¿Cómo y de dónde surgieron estas abstracciones matemáticas? ¿Qué base real, qué valor y aplicación reales tienen? ¿Cuál es su relación con la realidad? ¿Cómo se definen y cómo se estudian en matemáticas? ¿Cuál es el significado de las ideas generales de la geometría contemporánea en matemáticas?

Estas cuestiones serán contestadas en el presente capítulo. No haremos una exposición de la teoría de espacios matemáticos abstractos en sí; ello requeriría una mayor extensión y una atención mucho mayor al aparato estrictamente matemático. Nuestra tarea

consiste en arrojar luz sobre la esencia de las nuevas ideas de la geometría, es decir, contestar las preguntas planteadas aquí, y esto se puede hacer sin complicadas demostraciones y prolijas fórmulas.

La historia del problema que nos hemos planteado comienza con los *Elementos* de Euclides y el axioma —o, como también se dice, el postulado— de las rectas paralelas.

§1. Historia del postulado de Euclides

En sus *Elementos*, Euclides formula las premisas fundamentales de la geometría en forma de los llamados postulados y axiomas. Entre ellos figura el Postulado V (en otras ediciones de los *Elementos* es el Axioma XI) que se suele enunciar del siguiente modo: «Por un punto exterior a una recta se puede trazar una y sólo una paralela a dicha recta». Recordamos que una recta es paralela a otra si ambas están en un plano y no se cortan; aquí estamos pensando en rectas infinitas, no en segmentos finitos.

Es fácil probar que siempre se puede trazar al menos una paralela a una recta dada a desde un punto A exterior a ella.

Para ello tracemos desde A una perpendicular b a la recta dada a , y, pasando también por A , una recta c perpendicular a b (Fig. 1).

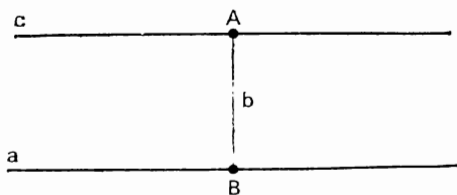


FIG. 1

La figura así obtenida es completamente simétrica respecto a b , ya que los ángulos que forma b con a y c en sus extremos son iguales. Por tanto, si doblamos el plano por b , las semirrectas de a , por un lado, y de c , por otro, se superponen. De aquí, que si a y b se cortaran a uno de los lados de AB también deberían cortarse al otro. Y ocurriría entonces que tendrían dos puntos en común; pero esto es imposible, pues, de acuerdo con una propiedad fundamental, por

dos puntos pasa una sola recta (de forma que dos rectas que tienen dos puntos comunes deben coincidir).

Así, de las propiedades básicas de la recta y del movimiento de una figura (ya que lo hecho anteriormente es una rotación de un semiplano alrededor de una recta contenida en él) se deduce que por un punto exterior a una recta se puede trazar al menos una paralela a dicha recta. El postulado de Euclides complementa este resultado afirmando que esta paralela tiene que ser única y que no existe otra.

Entre los demás postulados (axiomas) de la geometría, éste ocupa un lugar en cierto modo especial. La formulación de Euclides es bastante complicada, pero incluso en la forma citada aquí encierra una dificultad. Esta dificultad es ya inherente a la misma definición de rectas paralelas, donde, al hablar de «recta», nos referimos a la recta en su totalidad. ¿Pero cómo convencernos de que dos rectas son paralelas? Para ello tendríamos que prolongarlas por ambos lados «hasta el infinito» y cerciorarnos de que no se cortan nunca *en toda su infinita extensión*. Evidentemente, esta noción tiene sus dificultades. Y ésta fue, al parecer, la razón por la cual el postulado de las paralelas ocupaba incluso en la mente de Euclides un lugar especial: en los *Elementos* este postulado aparece por primera vez en la proposición 29 prescindiendo de él en las 28 anteriores. En vista de la complicada naturaleza del postulado es muy natural el deseo de prescindir de él; ya en la Antigüedad se hicieron intentos de cambiar la definición de rectas paralelas, de modificar la definición del postulado mismo, o mejor dicho, de deducirlo, como teorema, a partir del resto de los axiomas y de los conceptos básicos de la geometría.

Así, la teoría de las rectas paralelas, basada en el Postulado V, se convirtió en objeto de comentarios y críticas en los trabajos de muchos geómetras desde los tiempos antiguos. El objeto principal de estas investigaciones fue suprimir el Quinto Postulado y deducirlo como teorema de otras proposiciones básicas de la geometría.

La tarea atrajo a muchos geómetras: el griego Proclo (siglo V antes de C.), que escribió un comentario sobre Euclides; el persa Nasir ed Din et Tusi¹ (siglo XIII); el inglés Wallis (1616-1703); el italiano Saccheri (1667-1733); el filósofo y matemático alemán Lambert (1728-1777); el francés Legendre (1752-1833) y muchos otros; durante los dos milenios largos desde la aparición de los *Elementos* de Euclides, todos ellos se superaron en sutileza e ingenio geométrico para probar el Quinto Postulado.

Sin embargo, el resultado de estos intentos fue invariablemente negativo. Una y otra vez se demostró que el autor de turno se había apoyado en alguna proposición, tomada como obvia, que no se deducía necesariamente de las demás premisas de la geometría. Es decir, había sustituido el Quinto Postulado por otra proposición de la cual, en efecto, se seguía este postulado, pero que requería a su vez una demostración².

Saccheri y Lambert penetraron más profundamente que los demás en el problema. Saccheri fue el primero en dar una demostración del Postulado V por reducción al absurdo; es decir, tomó como punto de partida la aserción opuesta y, desarrollando las consecuencias, esperaba llegar a una contradicción. Cuando llegó, en sus inferencias, a resultados que le parecieron inimaginables, creyó haber resuelto el problema. Pero estaba equivocado, ya que una contradicción en las ideas intuitivas no tiene por qué ser una contradicción lógica. Después de todo, el problema consistía en probar el postulado euclidiano sobre la base de otras proposiciones de la geometría, y no en convencerse a sí mismo una vez más de su verdad intuitiva. Intuitivamente, el postulado en sí es suficientemente convincente. Pero, repitémoslo, las conclusiones intuitivas y la necesidad lógica son dos cosas distintas.

Lambert demostró ser un pensador más profundo que Saccheri y sus predecesores. Emprendiendo el mismo camino no encontró contradicción lógica ni cayó en los mismos errores que los demás; tampoco proclamó haber probado el Quinto Postulado. Pero después de él, a comienzos del siglo XIX, Legendre «prueba» una vez más el Quinto Postulado, cayendo en la vieja equivocación: sustituye de nuevo el postulado por otras aserciones que requieren una demostración.

Así, a comienzos del siglo XIX el problema se encuentra en la misma situación que en la época de Euclides. Los esfuerzos resultaban vanos y el problema parecía que no iba a ceder nunca. Se trataba efectivamente de un profundo enigma de la geometría; un problema, cuya

² Se han dado un gran número de tales enunciados equivalentes al Quinto Postulado. He aquí algunos ejemplos: 1) Una paralela a una recta dada dista de ella una longitud constante (Proclo); 2) Existen triángulos semejantes (pero no iguales), es decir, triángulos cuyos ángulos son iguales pero de lados desiguales (Wallis); 3) Existe al menos un rectángulo, esto es, un cuadrilátero cuyos ángulos son rectos (Saccheri); 4) Una recta perpendicular a un lado de un ángulo agudo también corta al otro lado (Legendre); 5) La suma de los ángulos de un triángulo es igual a dos rectos (Legendre); 6) Existen triángulos de área arbitrariamente grande (Gauss). La lista podría prolongarse sin fin.

solubilidad no fue puesta en duda por los mejores geómetras, llevaba dos mil años sin encontrar solución.

La teoría de las rectas paralelas se convirtió a lo largo del siglo XIX en uno de los principales problemas de la geometría, atrayendo a muchos geómetras: Gauss, Lagrange, d'Alambert, Legendre, Wachter, Schweikart, Taurinus, Farkas Bolyai y otros.

Pero la solución del problema no aparecía. ¿Qué era entonces lo que pasaba? ¿Era debido a una falta de capacidad o quizá a que el problema estaba mal planteado? La cuestión comenzó a preocupar a algunos geómetras que destacaban por su profundidad de pensamiento. Gauss, el matemático alemán más famoso, se ocupó del problema a partir de 1792 y poco a poco comenzó a vislumbrar el planteamiento correcto. Finalmente decidió abandonar el Quinto Postulado, y a partir de 1813 desarrolló una serie de teoremas como consecuencias de la aserción opuesta. Un poco más tarde, el matemático alemán Schweikart, por entonces profesor de Derecho en Tarkov, y Taurinus siguieron el mismo camino. Pero ninguno de ellos llegó a dar la respuesta final del problema. Gauss ocultó cuidadosamente sus investigaciones, Schweikart se limitó a la correspondencia privada con Gauss, y sólo Taurinus publicó los elementos de una nueva geometría basada en la negación del Quinto Postulado. Sin embargo, él mismo desechó la posibilidad de esta nueva geometría. Así, ninguno resolvió el problema, y la cuestión de saber si el planteamiento era enteramente correcto quedó sin respuesta. La solución fue dada, por primera vez, por N. I. Lobachevski, joven profesor de la Universidad de Kazan: el 23 de febrero de 1826 leyó una memoria sobre la teoría de las paralelas en una sesión de la facultad de física y matemáticas, y en 1829 publicó su contenido en la revista de la Universidad de Kazan.

§2. La solución de Lobachevski

1. Las ideas de Lobachevski

La esencia de la solución de Lobachevski al problema del Quinto Postulado la expresó él mismo en su trabajo *Nuevos Elementos de Geometría* (1835), en los siguientes términos:

«Es bien sabido que, en geometría, la teoría de las rectas paralelas ha permanecido hasta ahora incompleta. Los inútiles esfuerzos realizados desde los tiempos de Euclides a lo largo de dos mil años me han inducido a sospechar que los conceptos no contienen la

verdad que queríamos probar, sino que, al igual que otras leyes físicas, solamente pueden ser verificados mediante experimentos, tales como observaciones astronómicas. Convencido por fin de la verdad de mi conjetura y considerando que este difícil problema está completamente resuelto, expuse mis argumentos en 1826.»

Examinemos lo que Lobachevski quiere decir en este pasaje, en el que se enfoca, digámoslo así, su nueva idea. Idea que no sólo dio la solución del problema planteado por el Quinto Postulado, sino que también revolucionó nuestra concepción de la geometría y de otras ramas de las matemáticas.

Ya en 1815, I. N. Lobachevski comenzó a trabajar en la teoría de las paralelas, tratando al principio, como los demás geómetras, de probar el Quinto Postulado. En 1823 percibió claramente que de todas las demostraciones «ninguna de ellas, sea del tipo que sea, merece llamarse una demostración matemática en el completo sentido de la palabra, y sólo se pueden tomar como aclaraciones»³. Se había dado ya cuenta de que «los conceptos en sí no contienen la verdad que intentábamos probar»; es decir, el Quinto Postulado no se puede deducir de los conceptos y proposiciones fundamentales de la geometría. ¿Cómo llegó a la conclusión de que esta deducción era imposible?

La respuesta es: prosiguiendo el camino en el que Saccheri y Lambert habían dado los primeros pasos. Como hipótesis tomó la aserción contraria al postulado de Euclides, a saber: «por un punto exterior a una recta se pueden trazar no una, sino al menos dos rectas paralelas a la recta dada». Tomemos esta proposición, condicionalmente, como axioma y, después de añadir las demás proposiciones de la geometría, desarrollemos las consecuencias. Si la aserción es incompatible con las demás proposiciones de la geometría llegaremos a una contradicción y probaremos así indirectamente el Quinto Postulado: la proposición opuesta conduce a una contradicción. Sin embargo, como tal contradicción no se detecta por ningún lado, llegamos a dos conclusiones que Lobachevski también alcanzó.

La primera conclusión es que el Quinto Postulado no se puede probar. La segunda, que sobre la base de la proposición opuesta se puede desarrollar una cadena de consecuencias, es decir, teoremas, que no contienen contradicción alguna. Estas consecuencias forman por derecho propio una teoría no contradictoria, lógicamente posible,

³ Así escribía Lobachevski en 1823 en su curso de «Geometría», que no se publicó en vida del autor; este curso de «Geometría» fue editado por primera vez en 1910.

que puede considerarse como una nueva geometría no euclidiana. Lobachevski prudentemente la llamó «imaginaria», ya que no encontró un modelo real para ella. Pero vio claramente su posibilidad lógica. Y al expresar y defender su fuerte convicción, Lobachevski exhibió la verdadera grandeza de un genio que defiende su convicción sin vacilaciones y no la oculta de la opinión pública por miedo a la incompreensión y a la crítica.

Así, pues, las dos conclusiones a las que llegó Lobachevski fueron: primero, que el Quinto Postulado no se puede probar, y segundo, que es posible desarrollar sobre la base de un axioma contrario una nueva geometría completamente lógica, tan rica y perfecta como la de Euclides, a pesar de que sus resultados estén en desacuerdo con la imagen intuitiva del espacio. Lobachevski desarrolló, de hecho, la nueva geometría que lleva su nombre. Todo ello implicaba un resultado general de enorme importancia: *Existe más de una geometría lógicamente concebible*. El significado de este resultado será discutido más tarde en toda su extensión; en él está contenido una parte no pequeña de la solución a los problemas concernientes a los espacios matemáticos abstractos planteados al comienzo del capítulo.

Pero volvamos a la afirmación de Lobachevski citada al principio. Dice que la verdad geométrica, como otras leyes físicas, sólo puede verificarse mediante experimentos. Esto significa que debemos interpretar la verdad como una correspondencia entre los conceptos abstractos y la realidad. Esta correspondencia solamente se puede establecer por experimentos, de forma que la verificación de un resultado u otro requiere investigaciones experimentales, siendo insuficiente para ello la mera inferencia lógica. Aunque la geometría de Euclides refleja muy exactamente las propiedades del espacio, puede ocurrir que otras investigaciones revelen que dicha geometría, como teoría de las propiedades del espacio real, sea sólo aproximadamente verdadera. La geometría como ciencia del espacio real (y no solamente como sistema lógico) debería entonces ser modificada y reelaborada de acuerdo con los nuevos datos experimentales.

Esta brillante idea de Lobachevski ha sido completamente corroborada por una nueva rama de la física, la teoría de la relatividad.

El mismo Lobachevski realizó cálculos basados en observaciones astronómicas con objeto de comprobar la exactitud de la geometría de Euclides. Los cálculos confirmaron que era verdadera dentro de los límites de exactitud disponibles. La situación es ahora distinta, aunque hay que decir en seguida que la geometría de Lobachevski no ha resultado ser más correcta en su aplicación al espacio, cuyas

propiedades son diferentes y más complicadas. Esto no obstante, la geometría de Lobachevski está bien fundamentada y es aplicable en otros campos, de los cuales hablaremos luego con mayor detalle.

Hay que señalar que Lobachevski no consideró en absoluto su geometría como un mero esquema lógico construido sobre unas premisas arbitrariamente elegidas. La tarea capital la veía él no en el análisis lógico de los fundamentos de la geometría, sino en su relación con la realidad. Puesto que la experiencia no puede dar una solución absolutamente exacta al problema de la verdad del postulado de Euclides, tiene sentido investigar aquellas posibilidades lógicas que vienen representadas por las premisas más fundamentales de la geometría. Esta investigación matemática ayuda a marcar un camino por el cual debe proceder el estudio físico de las propiedades del espacio real. Además, la geometría de Euclides es un caso límite de la de Lobachevski, incluyendo esta última mayores posibilidades. Desde este punto de vista, la restricción al postulado de Euclides hubiera sido un impedimento para el desarrollo de la teoría. La teoría debe ir más allá de las fronteras conocidas e investigar caminos que descubran nuevos hechos y leyes. Una más profunda comprensión de los lazos entre las matemáticas y la realidad permite seleccionar, de entre las diversas posibilidades lógicas, aquellas que pueden tener más éxito en el estudio de la naturaleza. Si la geometría, después de Lobachevski, no hubiera desarrollado la doctrina matemática de las posibles propiedades del espacio, la física contemporánea no poseería los instrumentos matemáticos que permitieron formular y desarrollar la teoría de la relatividad.

Podemos resumir la solución de Lobachevski al problema del Quinto Postulado del siguiente modo:

1. El postulado no puede ser probado.
2. Añadiendo a las proposiciones básicas de la geometría el axioma opuesto se puede desarrollar una geometría extensa y lógicamente perfecta.
3. La verdad de los resultados de cualquier geometría lógicamente concebible, lo que atañe a sus aplicaciones al espacio real, sólo se puede verificar empíricamente. Una geometría lógicamente concebible debe ser desarrollada no sólo como un esquema lógico arbitrario, sino como una teoría que abra nuevos caminos y métodos para las teorías físicas.

La solución es totalmente distinta de la que buscaban los geómetras que trataron de probar el postulado euclidiano. Y esta solu-

ción era tan contraria a las ideas establecidas que no encontró comprensión entre los matemáticos de la época. Para ellos era algo demasiado nuevo y radical. Lobachevski cortó, por decirlo así, el nudo gordiano de la teoría de las paralelas, en lugar de intentar desenredarlo como otros trataban de hacer.

2. *Otros geómetras y filósofos*

Casi al mismo tiempo que Lobachevski, el geómetra húngaro Janos Bolyai (1802-1860) descubrió también la imposibilidad de probar el Quinto Postulado y la posibilidad de una geometría no euclidiana; publicó sus resultados en el apéndice a un tratado de geometría de su padre Farkas Bolyai en 1832. Ya antes, el padre había enviado a Gauss los resultados de su hijo y recibió una alentadora respuesta en la que Gauss decía haber llegado hacía mucho tiempo al mismo resultado. Sin embargo, Gauss se abstuvo de publicar nada. En una de sus cartas explicó que tenía miedo a ser mal interpretado.

En el terreno científico sucede siempre que, una vez llegado el momento para ciertos resultados, éstos suelen ser obtenidos casi simultánea e independientemente por distintos investigadores. El cálculo diferencial e integral fue desarrollado simultáneamente por Newton y Leibnitz; las ideas de Darwin fueron descubiertas independientemente por Wallace; los elementos de la teoría de la relatividad fueron hallados simultáneamente por Einstein y Poincaré, y podríamos citar muchos ejemplos más. Esto muestra que la ciencia se desarrolla inevitablemente mediante la solución de problemas para los que está madura, y no mediante descubrimientos accidentales y conjeturas. Así fue como el descubrimiento de la posibilidad de una geometría no euclidiana fue obra simultánea de varios geómetras: Lobachevski, Bolyai, Gauss, Schweikart y Taurinus.

Sin embargo, también ocurre constantemente en la ciencia que no todos los estudiosos que llegan a un mismo resultado juegan el mismo papel en su establecimiento ni tienen igual parte en el servicio prestado. La prioridad es importante, pero también lo es la claridad y profundidad de los resultados y la coherencia y solidez de los razonamientos. Ni Schweikart ni Taurinus estaban convencidos de la posibilidad e importancia de la nueva geometría, y esto fue decisivo en nuestro caso, tanto más cuanto que los resultados parciales habían sido ya obtenidos por Saccheri y Lambert. Gauss, aunque al parecer sí tenía esta convicción, no fue lo suficientemente resuelto para exponer a la luz pública sus resultados.

Bolyai no tuvo indecisiones, pero tampoco desarrolló la nueva

geometría tan profundamente como Lobachevski. Fue, pues, Lobachevski el primero que expresó abiertamente las nuevas ideas, verbalmente en 1826 y en letra impresa en 1829, continuando luego desarrollándolas y propagándolas mediante un gran número de escritos que culminaron en la *Pangeometría* de 1855, que dictó ya de viejo y ciego en sus últimos años, conservando aún su fortaleza mental y la confianza en su trabajo. Y por eso la nueva geometría lleva su nombre.

§3. *La geometría de Lobachevski*

1. *Algunos resultados sorprendentes*

Como ya hemos dicho, Lobachevski tomó como punto de partida la aserción opuesta al Quinto Postulado: en un plano dado, y por un punto exterior a una recta, se pueden trazar al menos dos rectas que no cortan a aquélla. De aquí dedujo numerosas consecuencias, que constituyen la nueva geometría. Así, pues, esta geometría fue construida como una colección de teoremas que se pueden probar a partir del postulado, junto con otros supuestos⁴ básicos de la geometría euclidiana, o, como decía Lobachevski, de la geometría «usual».

Entre sus deducciones Lobachevski obtuvo todos los resultados paralelos a los de la geometría elemental «usual», es decir, llegó a establecer la trigonometría no euclidiana, resolución de triángulos y cálculo de áreas y volúmenes. No podemos seguir aquí toda la cadena de deducciones hechas por Lobachevski, y no porque sean demasiado complicadas, sino simplemente por falta de espacio. Después de todo, incluso un curso escolar de geometría «usual» elemental es más bien largo, y las deducciones de Lobachevski no son ni más simples ni más breves que las deducciones «usuales». Mencionaremos, por tanto, sólo los resultados notables, remitiendo al lector interesado en un estudio más profundo a la literatura especializada. Más adelante haremos una interpretación simple de la geometría no euclidiana en el mundo actual.

Comencemos con la teoría de las líneas paralelas. Supongamos que nos dan una línea a y un punto A exterior a ella. Por A trazamos la perpendicular AB a la línea a . En virtud del supuesto fundamental existen al menos dos líneas que pasan por A y no cortan a la línea

⁴ Estas «restantes» proposiciones de la geometría se formularán con exactitud en §5.

dada a . Entonces ninguna de las líneas comprendidas entre ambas cortará tampoco a a . Ciertamente es que en la figura 2 las líneas b y b' , si se prolongan suficientemente, cortan a a , en contra de lo supuesto por Lobachevski. Pero esto no debe sorprendernos. Lobachevski, naturalmente, no razonó a base de figuras como las que se trazan en el plano ordinario, sino que desarrolló las consecuencias *lógicas* de sus suposiciones (que son contrarias a lo que estamos acostum-

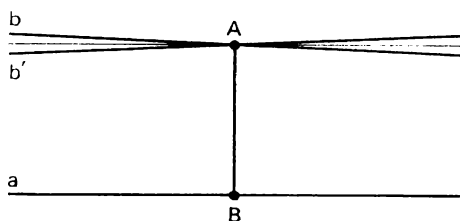


FIG. 2

brados a ver en los diagramas). Las figuras y diagramas juegan un papel estrictamente auxiliar; en ellas no se pueden expresar las situaciones de ninguna geometría no euclidiana, ya que dichas figuras representan rectas ordinarias en el plano ordinario, y este plano es completamente euclidiano dentro de los límites de exactitud de la figura.

Esta contradicción entre la posibilidad lógica y la representación visual fue un importante obstáculo para la comprensión de la geometría de Lobachevski. Pero, si estamos interesados en la geometría como teoría lógica, debemos atender al rigor lógico y no a la conformidad con las figuras usuales.

2. El ángulo de paralelismo

Volvamos otra vez a nuestra línea a y al punto A . Por A trazamos una semirrecta x que no corte a a (por ejemplo, la perpendicular a AB) y la giramos con centro en A de forma que el ángulo ϕ entre AB y x decrezca, pero sin llegar a cortar a a . Entonces la semirrecta x alcanza una posición límite correspondiente al valor mínimo de ϕ . Esta semirrecta límite c tampoco corta a a .

Para probarlo supongamos que c corta a a en algún punto X (Fig. 3); podemos entonces tomar un punto X' a la derecha de X y trazar la semirrecta AX' , que corta a a y al mismo tiempo forma con AB

un ángulo mayor. Pero esto es imposible, ya que, por construcción, toda semilínea que forma con AB un ángulo mayor que c no corta a a .

Por consiguiente, c no corta a a y es además la más extrema de todas las semilíneas que pasando por A no cortan a a .

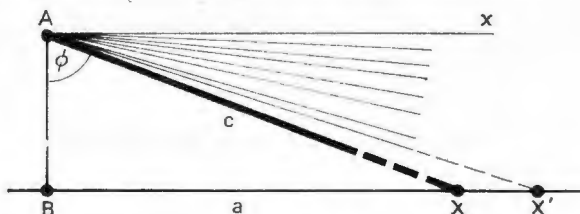


FIG. 3

Es evidente que, por simetría, podemos trazar en el lado opuesto otra semilínea c' que no corte a a y que sea la más extrema de entre todas las que, como ella, tampoco cortan a a . Si c y c' fueran continuación una de otra, formarían juntas una única línea $c + c'$. Esta sería entonces la única paralela a a que pasaría por A , ya que con el más mínimo giro, o bien c , o bien c' cortarían a a . Ahora bien, como se ha supuesto que la paralela no es única sino que al menos existen dos, las semilíneas c y c' no pueden ser prologación una de otra.

Hemos probado así el primer teorema de la geometría de Lobachevski.

Por un punto A exterior a una línea a se pueden trazar dos semilíneas c y c' que no cortan a dicha línea, y todas las semilíneas comprendidas en el ángulo que forman entre ellas cortan a a .

Si prolongamos las semilíneas c y c' obtenemos (Fig. 4) dos líneas

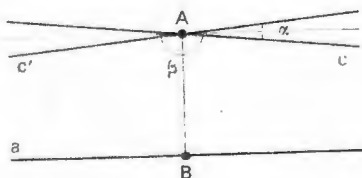


FIG. 4

que no cortan a a con la propiedad de que todas las líneas que pasan por A dentro del ángulo α formado por ellas no cortan a a , y las demás líneas en el ángulo β sí la cortan. Lobachevski llamó a estas dos líneas c y c' *paralelas a a* : c paralela a la derecha y c' paralela a la izquierda.

La mitad del ángulo β fue llamado por Lobachevski *ángulo de paralelismo*; es menor que un recto, ya que β es menor que dos rectos.

3. *Convergencia de líneas paralelas; la curva equidistante*

Investiguemos ahora cómo varía la distancia de a a un punto X de c cuando X se mueve a lo largo de c (Fig. 5). En la geometría euclidiana la distancia entre rectas paralelas es constante. Aquí, sin embargo, podemos darnos cuenta de que cuando X se mueve hacia la derecha la distancia a a disminuye (es decir, la longitud de la perpendicular XY decrece).

Tracemos la perpendicular A_1B_1 desde un punto A_1 a a . Por B_1 trazamos la perpendicular B_1A_2 a c (A_2 está a la derecha de A_1 , ya que γ es un ángulo agudo). Finalmente trazamos desde A_2 la perpendicular A_2B_2 a a . Mostremos que A_2B_2 es menor que A_1B_1 .

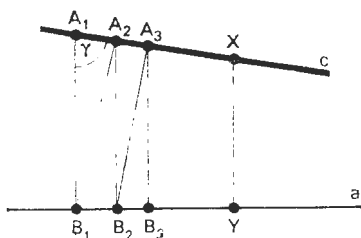


FIG. 5

El teorema que afirma que la perpendicular es más corta que cualquier oblicua sigue siendo válido en la geometría de Lobachevski, ya que su demostración (que se puede encontrar en cualquier libro escolar de geometría) no depende del concepto de rectas paralelas ni de las consecuencias derivadas de él. Pues bien, como la perpendicular es menor que la oblicua, B_1A_2 , que es perpendicular a c , es menor que A_1B_1 , y por la misma razón A_2B_2 es menor que B_1A_2 . Por tanto, A_2B_2 es menor que A_1B_1 .

Trazando desde B_2 la perpendicular B_2A_3 a c y repitiendo los razonamientos vemos que A_3B_3 es menor que A_2B_2 . Continuando la construcción, obtenemos una sucesión de perpendiculares cada vez más pequeñas, es decir, las distancias desde $A_1, A_2, \dots$, hasta a decrecen. Además, continuando este sencillo razonamiento se puede probar que, si un punto X'' de c está a la derecha de X' , la perpendicular $X''Y''$ es más pequeña que $X'Y'$. No insistimos en este punto. Los argumentos precedentes creemos que aclaran suficientemente la esencia de la cuestión y, por otra parte, no es nuestro propósito dar una demostración rigurosa.

Pero es notable que, como se puede probar, la distancia XY no sólo disminuye cuando X se desplaza hacia la derecha, sino que tiende a cero cuando X tiende hacia infinito. Es decir *las paralelas a y c convergen asintóticamente*. Se puede probar además que en el sentido opuesto la distancia entre ellas no sólo crece, sino que tiende a infinito.

En la geometría euclidiana la distancia entre dos rectas paralelas permanece constante. En la geometría de Lobachevski, en general, no existen tales pares de líneas, ya que una recta siempre diverge de otra dada, bien por un lado, bien por ambos. Así, pues, una línea que guarde una distancia constante respecto a otra dada no puede ser recta; es una curva llamada *equidistante*.

Las conclusiones de Lobachevski son efectivamente sorprendentes y desde luego no compatibles con la representación visual acostumbrada. Pero, como ya hemos dicho, una discrepancia tal no puede ser un argumento en contra de la geometría de Lobachevski considerada como teoría abstracta, desarrollada lógicamente a partir de las premisas supuestas.

4. La magnitud del ángulo de paralelismo

Vamos a estudiar ahora el ángulo de paralelismo, es decir, el ángulo γ que la línea c paralela a la línea dada a forma con la perpendicular CA (Fig. 6). Demostremos que este ángulo es tanto más

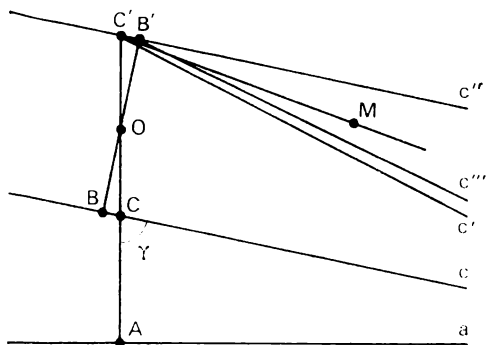


FIG. 6

pequeño cuanto más lejano está C de a . Para ello comencemos probando lo siguiente. Si dos líneas b y b' forman ángulos iguales α y α' con una secante BB' , entonces tienen una perpendicular común (Fig. 7).

Para demostrarlo trazaremos por el punto medio O de BB' la línea CC' perpendicular a b . Obtenemos dos triángulos OBC y $OB'C'$. Los lados OB y OB' son iguales por construcción. Los ángulos en el vértice común O son iguales por ser opuestos por el vértice.

El ángulo α'' es igual a α' por la misma razón. Pero α' es igual a α por hipótesis. Por tanto α es igual a α'' . Así, en los triángulos OBC y $OB'C'$ los lados OB y OB' y sus ángulos adyacentes son iguales. Pero entonces, por un conocido teorema, los dos triángulos son iguales y, en particular, los ángulos en C y C' también. Ahora bien, el ángulo en C es recto puesto que la línea CC' es perpendicular a b por construcción. Por tanto el ángulo en C' es también recto; es decir, CC' es también perpendicular a b' . Así, pues, el segmento CC' es una perpendicular común a b y b' . Esto prueba la existencia de una perpendicular común.

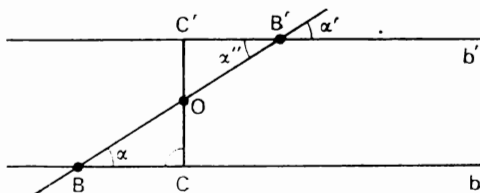


FIG. 7

Probemos ahora que el ángulo de paralelismo decrece cuando aumenta la distancia del punto a la recta. Esto es, si el punto C' está más alejado de a que C , entonces, como ocurre en la figura 6, la paralela c' que pasa por C' forma con la perpendicular $C'A$ un ángulo más pequeño que la paralela c que pasa por C .

Para la demostración trazamos por C' una línea c'' que forme con $C'A$ el mismo ángulo que la paralela c . Entonces c y c'' forman con CC' ángulos iguales. Por tanto, como acabamos de probar, tienen una perpendicular común BB' . Podemos trazar entonces por B' una línea c''' paralela a c , que debe formar con la perpendicular un ángulo menor que un recto, pues como ya sabemos una paralela forma con la perpendicular un ángulo menor que un recto. Tomamos ahora un punto arbitrario M en el ángulo formado por c'' y c''' y trazamos la línea $C'M$. Esta línea se encuentra en el ángulo formado por c' y c'' y no corta a a . *A fortiori*, no puede cortar a c . Pero forma con AC' un ángulo menor que c'' , es decir, menor que γ . Entonces, *a fortiori*, la paralela c' forma un ángulo aún menor porque es la más extrema de las líneas que pasando por C' no cortan a a . Por tanto, c' forma con $C'A$ un ángulo menor que el que forma c , y esto significa que el ángulo de paralelismo disminuye al pasar a un punto C' más lejano; esto es lo que queríamos demostrar.

Hemos demostrado entonces que el ángulo de paralelismo decrece al aumentar la distancia de C a a . Se puede demostrar incluso más: *si el punto C se aleja hasta el infinito, el ángulo tiende hacia cero*. Esto es, para una distancia suficientemente grande de a , la paralela a ella forma con la perpendicular a a un ángulo arbitrariamente pequeño⁵. En otros términos, si en un punto muy lejano de a inclinamos la perpendicular a a un ángulo muy pequeño, esta línea «inclinada» ya no cortará a a . Este resultado de la geometría de Lobachevski produce impresión desconcertante. Más adelante encontraremos otros resultados no menos asombrosos.

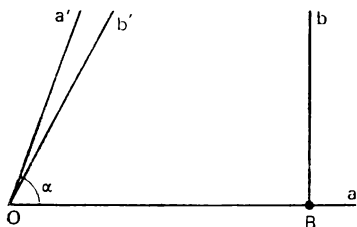


FIG. 8

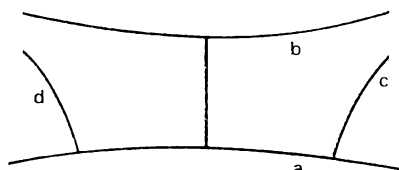


FIG. 9

Tomemos, por ejemplo, el ángulo agudo α formado por dos semilíneas a y a' . Levantamos la perpendicular b a a en un punto suficientemente alejado del vértice O del ángulo α de forma que el ángulo de paralelismo correspondiente a la distancia OB (Fig. 8) elegida sea menor que α . Como el ángulo α es mayor que el ángulo de paralelismo, la línea b' que pasa por O y es paralela a b , forma con a un ángulo más pequeño. Pero b' no corta a b y por consiguiente a' tampoco corta a b . Hemos demostrado así que la perpendicular a uno de los lados de un ángulo agudo, levantada suficientemente lejos del vértice, no corta al otro lado.

5. Otros teoremas notables

Hemos expuesto los resultados precedentes con dos fines. En primer lugar, y este es el punto más importante, queríamos mostrar

⁵ Si h es la longitud de la perpendicular, y γ , el ángulo de paralelismo, entonces, como mostró Lobachevski, $\text{tg } \gamma/2 = e^{-h/k}$, donde k es una constante que depende de la unidad de longitud y e es la base de los logaritmos naturales. Evidentemente tanto $e^{-h/k}$ como γ tienden a cero cuando h tiende a infinito.

por medio de ejemplos sencillos cómo los teoremas de Lobachevski se pueden obtener, de hecho, de las premisas adoptadas. Se trata de un ejemplo muy simple de la forma en que los matemáticos obtienen resultados en geometría abstracta y de cómo se pueden obtener conclusiones completamente desligadas de la representación visual acostumbrada. En segundo lugar, queríamos mostrar que en la geometría de Lobachevski se obtienen resultados peculiares. Demos algunos ejemplos más.

Dos líneas en un plano de Lobachevski o se cortan o son paralelas en el sentido de Lobachevski, y en ese caso convergen asintóticamente por uno de los lados y por el otro divergen infinitamente, o, si no, tienen una perpendicular común y divergen infinitamente por ambos lados.

Si las líneas a y b tienen una perpendicular común (Fig. 9) se pueden trazar dos perpendiculares c y d a a que son paralelas (en el sentido de Lobachevski) a b y tal que toda la línea b está contenida en la franja comprendida entre c y d .

El límite de un círculo cuando el radio crece infinitamente no es una recta, sino una cierta curva llamada *círculo límite*. Por tres puntos no alineados no siempre es posible construir un círculo, pero siempre se puede hacer pasar por ellos o un círculo, o un círculo límite, o bien una equidistante (es decir, una línea formada por los puntos que equidistan de una cierta recta).

La suma de los ángulos de un triángulo es siempre menor que dos rectos. Aumentando un triángulo dado, de forma que sus tres lados crezcan indefinidamente, sus tres ángulos tienden a cero.

No existen triángulos de área arbitrariamente grande.

Dos triángulos son iguales cuando sus ángulos son iguales.

La longitud l de una circunferencia no es proporcional al radio r sino que crece más rápidamente que él (a través de una ley exponencial). Se tiene la siguiente fórmula:

$$l = \pi k (e^{r/k} - e^{-r/k}), \quad (1)$$

donde k es una constante que depende de la unidad de longitud. Teniendo en cuenta que

$$e^{r/k} = 1 + \frac{r}{k} + \frac{1}{2} \left(\frac{r}{k}\right)^2 + \cdots, \quad e^{-r/k} = 1 - \frac{r}{k} + \frac{1}{2} \left(\frac{r}{k}\right)^2 - \cdots,$$

de (1) obtenemos:

$$l = 2\pi r \left(1 + \frac{1}{6} \frac{r^2}{k^2} + \cdots\right) \quad (2)$$

Sólo para fracciones r/k pequeñas es verdad, con suficiente aproximación, que $l = 2\pi r$.

Todas estas conclusiones son consecuencias lógicas de las premisas iniciales: «el axioma de Lobachevski» junto con las proposiciones básicas de la geometría «usual».

6. La geometría de Lobachevski comparada con la de Euclides

Una propiedad extremadamente importante de la geometría de Lobachevski consiste en el hecho de que para dominios pequeños difiere poco de la geometría de Euclides; cuanto más pequeño es el dominio, menor es la diferencia. Así, para triángulos suficientemente pequeños las relaciones entre lados y ángulos se expresan, con un suficiente grado de aproximación, mediante las fórmulas de la trigonometría ordinaria, y esto más aproximadamente cuanto más pequeño sea el triángulo.

La fórmula (2) muestra que para radios pequeños la longitud de una circunferencia es proporcional al radio con una buena aproximación. Análogamente, la suma de los ángulos de un triángulo difiere muy poco de dos rectos, etc.

En la fórmula que da la longitud de la circunferencia entra una constante k que depende de la unidad de longitud. Si el radio es pequeño en comparación con k , es decir, si r/k es pequeño, entonces está claro por la fórmula (2) que la longitud l difiere poco de $2\pi r$. De modo general, cuanto más pequeña es la relación de las dimensiones de una figura a la constante k , tanto más se aproximan las propiedades de una figura a las de la figura correspondiente en la geometría de Euclides ⁶.

⁶ Por ejemplo, si a, b, c son los catetos y la hipotenusa de un triángulo rectángulo, entonces, en lugar del teorema de Pitágoras, se verifica la siguiente relación:

$$2(e^{c/k} + e^{-c/k}) = (e^{a/k} + e^{-a/k})(e^{b/k} + e^{-b/k})$$

Desarrollando en serie, obtenemos

$$c^2 + \frac{c^4}{12k^2} + \dots = a^2 + b^2 + \frac{a^4 + 6a^2b^2 + b^4}{12k^2} + \dots,$$

de modo que para l grande tenemos el teorema de Pitágoras $c^2 = a^2 + b^2$. Además, de la fórmula de Lobachevski, para el ángulo de paralelismo γ , $\operatorname{tg} \gamma = e^{-h/k}$ (véase nota 5), se deduce que, si h/k es pequeño, es decir, si las paralelas están próximas, entonces $\operatorname{tg} \gamma/2 = e^{-h/k} \approx 1$ y $\gamma = 90^\circ$. Así, pues, para distancias pequeñas, las paralelas en un plano de Lobachevski difieren poco de las paralelas euclidianas.

Una medida de la diferencia entre las propiedades de una figura en la geometría de Lobachevski y las de la correspondiente en la de Euclides, es la relación r/k , donde r mide las dimensiones de la figura (el radio de un círculo, los lados de un triángulo, etc.).

Esto tiene una consecuencia importante.

Supongamos que estamos trabajando en el espacio real del mundo exterior y que medimos las distancias en kilómetros. Supongamos también que la constante k es muy grande, pongamos 10^{12} .

Entonces, por la fórmula (2), para un círculo de radio 100 km, la razón de su longitud al radio difiere de 2π en menos de 10^{-9} . Del mismo orden son las diferencias respecto a la geometría euclidiana para otras razones. Dentro de los límites de 1 kilómetro serían incluso de orden $1/k$, es decir, 10^{-12} , y dentro de los límites de un metro del orden de 10^{-15} , es decir, completamente despreciables. Tales diferencias respecto a la geometría euclidiana no podrían ser observadas, ya que las dimensiones de un átomo son cien veces mayores (del orden de 10^{-13} km). Sin embargo, la razón r/k podría ser no despreciable a escala astronómica.

Por consiguiente, Lobachevski también supuso que, aunque a la escala ordinaria la geometría de Euclides es verdadera con gran aproximación, las diferencias respecto a ella podrían ser detectadas mediante observaciones astronómicas. Como ya dijimos anteriormente, esta suposición ha sido corroborada, pero las insignificantes diferencias respecto a la geometría de Euclides que se han observado a escala astronómica resulta que son de índole más compleja.

Finalmente, los razonamientos expuestos anteriormente tienen otra importante consecuencia: puesto que las diferencias respecto a la geometría de Euclides disminuyen para valores crecientes de la constante k , en el límite —cuando k crece indefinidamente— la geometría de Lobachevski se convierte en la de Euclides. Esto es, *la geometría de Euclides es únicamente un caso límite de la geometría de Lobachevski*. Por tanto, si incorporamos este caso límite a la geometría de Lobachevski, ésta contiene a la de Euclides y se convierte, en este sentido, en una teoría más general. Por esta razón, Lobachevski llamó a su teoría «pangeometría», es decir, geometría universal. Tal relación entre las teorías aparece constantemente en el desarrollo de las matemáticas y de las ciencias naturales: una nueva teoría incluye la antigua como caso límite, de acuerdo con el progreso de nuestro conocimiento de casos particulares a casos más generales.

Sin embargo, todos los razonamientos y deducciones anteriores quedarían, por decirlo así, en un juego apenas inteligible de la mente

si no pudiéramos establecer un significado real relativamente simple de la geometría de Lobachevski dentro del sistema de los conceptos usuales de la geometría euclidiana. La solución de este problema no fue hallada por el propio Lobachevski, sino por sus sucesores, casi cuarenta años después de la aparición de su primer escrito. Esta solución se describe en la próxima sección.

§4. El significado real de la geometría de Lobachevski

1. La interpretación de Beltrami sobre la pseudoesfera

La primera interpretación intuitiva de la geometría de Lobachevski se dio en 1868, cuando el geómetra italiano Beltrami hizo notar que la geometría intrínseca de una cierta superficie, la pseudoesfera, coincidía con la geometría sobre parte del plano de Lobachevski. Recordemos que por geometría intrínseca sobre una superficie se entiende la colección de propiedades que se pueden determinar midiendo longitudes solamente sobre la superficie. En la figura 10, a la izquierda, está representada la curva llamada *tractriz*. Esta curva tiene la propiedad de que la longitud del segmento de tangente comprendido entre el punto de tangencia y el punto de corte con el eje Oy es cons-

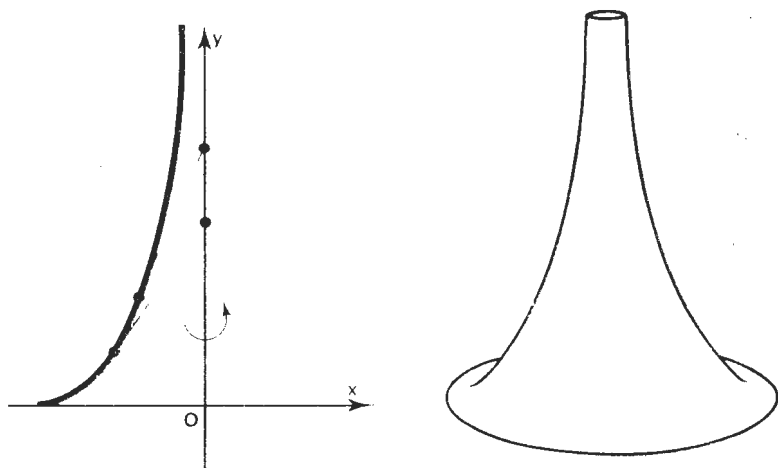


FIG. 10

tante. El eje Oy es una asíntota. Al girar la curva alrededor de su asíntota se engendra una superficie llamada *seudoesfera* representada en la figura 10 a la derecha.

La interpretación de Beltrami de la geometría de Lobachevski viene a decir que todas las relaciones geométricas sobre una parte del plano de Lobachevski coinciden con las relaciones geométricas sobre una parte conveniente de la seudoesfera adoptando los siguientes convenios. El papel de los segmentos de recta está desempeñado por las líneas de menor longitud sobre la superficie, las llamadas geodésicas. La distancia entre dos puntos se define como la longitud de la línea más corta que une a ambos sobre la superficie. Se dice que dos figuras son iguales si entre sus puntos se puede establecer una correspondencia, de forma que la distancia entre puntos correspondientes sea la misma. Un movimiento sobre la seudoesfera que conserve las dimensiones desde el punto de vista de la geometría intrínseca, aunque vaya acompañado de torsiones, representa un movimiento en el plano Lobachevski. Las longitudes, áreas y ángulos se miden, como de costumbre, sobre la superficie, y corresponden a longitudes, áreas y ángulos en la geometría de Lobachevski.

La interpretación de Beltrami muestra que, en estas condiciones, a cada proposición de la geometría de Lobachevski referida a una parte del plano corresponde un hecho inmediato en la geometría intrínseca sobre la seudoesfera. La geometría de Lobachevski tiene, en consecuencia, un perfecto significado real: no es más que una exposición abstracta de la geometría sobre la seudoesfera.

Debemos mencionar que treinta años antes del descubrimiento de Beltrami, la geometría intrínseca de la seudoesfera había sido ya estudiada por F. Minding, quien demostró de hecho las propiedades que demuestran la coincidencia entre las dos geometrías. Sin embargo, ni él ni nadie se percataron de ello hasta que las ideas de Lobachevski se divulgaron suficientemente. Beltrami sólo tuvo que comparar los resultados de Lobachevski y Minding para darse cuenta de la relación entre ambas.

El descubrimiento de Beltrami cambió por completo la actitud de los matemáticos hacia la geometría de Lobachevski; pasó de ser una cosa «ficticia» a ser algo real ⁷.

⁷ La historia del descubrimiento del significado real de la geometría de Lobachevski fue en realidad más complicada. En primer lugar, el propio Lobachevski tuvo a su alcance los medios necesarios para probar su carácter no contradictorio mediante un modelo de los llamados analíticos, pero no consiguió llevar a cabo la demostración. Esta fue dada mucho más tarde. En segundo

2. *La interpretación de Klein en el círculo y en la esfera*

Sin embargo, como hemos recalcado anteriormente, sobre la seudoesfera no se realiza más que una parte del plano de Lobachevski.⁸

Por tanto, el problema de dar una interpretación real de la geometría de Lobachevski sobre el plano completo y, sobre todo, de su geometría en el espacio, permanecía sin resolver. La solución fue dada más tarde en 1870 por el matemático alemán Klein. Expliquemos cuál fue su solución.

Tomemos, en el plano euclidiano ordinario, un círculo y consideremos solamente su interior, es decir, excluimos de nuestro estudio su circunferencia y el dominio exterior. Convenimos en llamar «plano» al interior del círculo, ya que, como veremos, juega el papel de plano de Lobachevski. Las cuerdas del círculo serán llamadas «rectas» y de acuerdo con el convenio anterior los puntos extremos serán excluidos por pertenecer a la circunferencia. Finalmente, un «movimiento» será cualquier transformación que aplique el círculo en sí mismo y que transforme rectas en rectas, es decir, que no distorsione las cuerdas. El ejemplo más sencillo de una transformación de este tipo es una rotación del círculo alrededor de su centro, pero

lugar, el matemático alemán Riemann puso a punto en 1854 una teoría (ver §10) en la que están contenidos los resultados de Beltrami, aunque Riemann no los expresara con claridad; su trabajo no fue comprendido y no se publicó hasta después de su muerte en 1868, el mismo año en que apareció el trabajo de Beltrami. Toda la historia de la geometría de Lobachevski, desde los primeros intentos de probar el postulado de Euclides hasta la completa aclaración del significado de la geometría no euclidiana, es altamente instructiva porque muestra cómo, a menudo, es necesario recorrer tortuosos e intrincados caminos para descubrir una verdad que después resulta simple e inteligible.

⁸ La seudoesfera tiene en todos sus puntos una curvatura gaussiana negativa constante. Todas las superficies de curvatura negativa constante tienen (al menos localmente) la misma geometría intrínseca y pueden servir por tanto como modelos de la geometría de Lobachevski. Sin embargo, como probó Hilbert en 1901, ninguna de estas superficies se puede extender infinitamente en todas las direcciones sin tropezar con singularidades, y en consecuencia no sirven para representar la geometría de Lobachevski en *todo* el plano. Por otra parte, el joven matemático holandés Kuiper demostró en 1955 que existen superficies lisas que representan —en el sentido de su geometría intrínseca— todo el plano de Lobachevski, pero que tales superficies no pueden ser deformadas continuamente y no tienen una curvatura definida.

Hagamos notar también lo siguiente. Cuando la geometría de Lobachevski se representa sobre una superficie de curvatura negativa constante K , la constante k que aparece en las fórmulas de la sección precedente adquiere un significado muy simple: $k^2 = -1/K$.

existen muchas más. Que estas transformaciones existen será demostrado en lo que sigue.

Introduciendo estos convenios de nomenclatura, los teoremas de la geometría ordinaria dentro del círculo se transforman en los teoremas de la geometría de Lobachevski. E inversamente, cada teorema de la geometría de Lobachevski se interpreta como un hecho de la geometría ordinaria dentro del círculo.

Por ejemplo, el axioma de Lobachevski dice que por un punto exterior a una recta se pueden trazar al menos dos rectas que no corten a la primera. Traslademos este axioma al lenguaje de la geometría ordinaria mediante nuestros convenios, es decir, reemplazando las rectas por cuerdas. Obtenemos entonces el siguiente enunciado: por un punto interior al círculo y exterior a una cuerda se pueden trazar al menos dos cuerdas que no cortan a la primera. Que esto es cierto resulta obvio por la figura 11. Por tanto, el axioma de Lobachevski se satisface en nuestro modelo.

Recordemos además que en la geometría de Lobachevski, entre todas las rectas que pasando por un punto no cortan a una recta dada existen dos rectas extremo. Esto significa que entre todas las cuerdas que pasan por un punto dado A y no cortan a una cuerda dada BC existen dos cuerdas extremo. Y, efectivamente, dichas cuerdas son las que pasan por B y C . No tienen puntos comunes con BC , ya que excluimos los puntos de la circunferencia. Así, pues, este teorema de Lobachevski se cumple aquí.

Para poder seguir traduciendo los teoremas de Lobachevski al lenguaje de la geometría ordinaria dentro del círculo es necesario explicar cómo se miden en el círculo los segmentos y los ángulos de forma que las medidas correspondan a la geometría de Lobachevski. Desde luego, el proceso de medida no puede ser el mismo que el usual, ya que una cuerda, en el sentido ordinario, tiene una longitud finita, mientras que la recta que la cuerda representa es infinita. Esto se puede considerar quizá como una contradicción, pero veremos que no hay tal.

Recordemos que en la geometría ordinaria la longitud de segmentos se mide como sigue. Tomamos un segmento AB como longitud unidad y determinamos la longitud de otro segmento XY por comparación con AB . Para ello colocamos el segmento AB a lo largo de XY . Si queda una parte de XY menor que AB dividimos éste en 10 partes iguales (iguales en el sentido de que cada una de obtiene de la anterior por traslación); estas partes se colocan encima de la porción restante de XY ; si fuera necesario dividimos AB en 100 partes, etc.

Así, la longitud de XY se expresa mediante una fracción decimal que puede tener incluso infinitas cifras. En consecuencia, las longitudes se miden mediante un movimiento de cierta parte del segmento tomado como unidad; es decir, la medida se basa en el movimiento. Y una vez definidos los movimientos (que en nuestro caso se definen como transformaciones del círculo en sí mismo que llevan rectas sobre rectas), sabemos qué segmentos hay que considerar como iguales y cuál es su longitud (cómo debemos medirla). Los términos en que se define el movimiento contienen ya, aunque de forma implícita, la regla para medir longitudes. Los ángulos se miden exactamente de la misma forma, por comparación con el ángulo tomado como unidad. Así, la regla para medir ángulos está también contenida en la definición de movimiento.

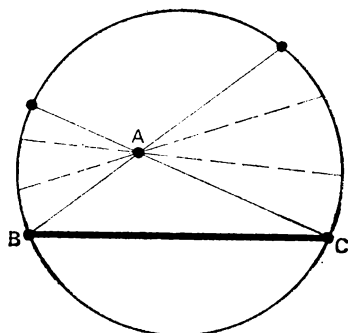


FIG. 11

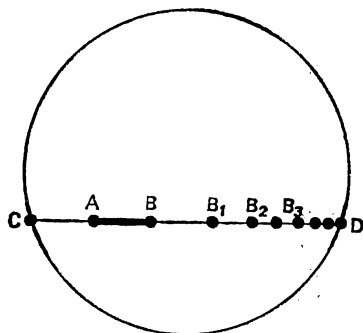


FIG. 12

Las correspondientes reglas para medir longitudes y ángulos en la geometría de Lobachevski son muy sencillas, aunque difieren esencialmente de las reglas ordinarias. No las estudiaremos aquí porque no son un punto de principal trascendencia para nuestros razonamientos ⁹.

Por la regla de medir longitudes resulta que una cuerda tiene longitud infinita. Y esto se debe a que, si, mediante una transformación que hayamos tomado como movimiento, transformamos el segmento AB en el BB_1 , éste en el B_1B_2 , etc., los segmentos B_kB_{k+1} serán

⁹ La regla para medir longitudes es la siguiente. Sea el segmento AB sobre la cuerda CD (Fig. 12). Midiendo segmentos de la forma habitual formamos la razón doble $CB/CA : DB/DA$. Su logaritmo se toma como la longitud de AB .

cada vez más pequeños en el sentido usual (aunque iguales en el sentido de nuestro modelo para la geometría de Lobachevski; figura 12). Los puntos $B_1, B_2, \dots, B_k, \dots$ se acumulan en el extremo de la cuerda. Pero para nosotros las cuerdas no tienen punto extremo; el extremo se excluye por convenio, y en este sentido se encuentra en «el infinito». En el sentido de la geometría de Lobachevski los puntos $B_1, B_2, \dots$ no se acumulan en ninguna parte, sino que tienden al infinito. Por medio de las transformaciones que hemos tomado como movimientos no podemos alcanzar los puntos de la circunferencia desde dentro del círculo sin más que colocar segmentos iguales uno a continuación de otro.

Para comprender mejor cómo se transportan los segmentos en este modelo, consideremos la transformación que juega el papel de una traslación a lo largo de una recta.

Introducimos un sistema de coordenadas rectangulares en el plano y como origen de coordenadas tomamos el centro del círculo. Supongamos, para fijar ideas, que el círculo tiene de radio la unidad, de forma que su circunferencia tiene de ecuación $x^2 + y^2 = 1$ y los puntos del interior satisfacen la desigualdad $x^2 + y^2 < 1$.

Consideremos la transformación dada por las fórmulas

$$x' = \frac{x + a}{1 + ax}, \quad y' = \frac{y \sqrt{1 - a^2}}{1 + ax}, \quad (3)$$

donde x' e y' son las coordenadas del punto en que se transforma el punto de coordenadas x e y . Aquí a es un número arbitrario de valor absoluto menor que 1.

Despejando x e y de (3), obtenemos, como es fácil de comprobar, para las expresiones inversas de x e y en función de x' e y'

$$x = \frac{x' - a}{1 - ax'}, \quad y = \frac{y' \sqrt{1 - a^2}}{1 - ax'}. \quad (4)$$

La transformación (3) satisface las dos condiciones para ser un «movimiento» en nuestro modelo: 1) transforma el círculo en sí mismo; 2) transforma rectas en rectas.

Para probar la primera propiedad es preciso observar que la relación $x^2 + y^2 \leq 1$ implica la correspondiente relación $x'^2 + y'^2 \leq 1$, y viceversa. Mostremos, por ejemplo, que cuando $x^2 + y^2 = 1$ necesariamente $x'^2 + y'^2 = 1$, es decir, que los puntos de la circunferencia se transforman en puntos de la circunferencia.

Calculemos $x'^2 + y'^2 = 1$ haciendo uso de (3) y teniendo en cuenta que $x^2 + y^2 = 1$, es decir, $y^2 = 1 - x^2$,

$$\begin{aligned} x'^2 + y'^2 &= \frac{(x + a)^2 + y^2(1 - a^2)}{(1 + ax)^2} = \frac{(x + a)^2 + (1 - x^2)(1 - a^2)}{(1 + ax)^2} = \\ &= \frac{x^2 + 2ax + a^2 + 1 - x^2 - a^2 + a^2x^2}{(1 + ax)^2} = \frac{1 + 2ax + a^2x^2}{1 + 2ax + a^2x^2} = \\ &= 1. \end{aligned}$$

Por tanto cuando $x^2 + y^2 = 1$ también $x'^2 + y'^2 = 1$. El resto de los casos se verifican de forma similar.

La segunda propiedad de (3) se establece muy fácilmente. Sabemos que una recta se representa mediante una ecuación lineal; e, inversamente, una ecuación lineal representa una recta. Supongamos dada una recta de ecuación

$$Ax + By + C = 0. \quad (5)$$

Después de la transformación (4) obtenemos

$$A \frac{x' - a}{1 - ax'} + B \frac{y' \sqrt{1 - a^2}}{1 - ax'} + C = 0,$$

o, reduciendo a común denominador,

$$(A - aC)x' + B\sqrt{1 - a^2}y' + (C - aA) = 0.$$

Esta ecuación es lineal y por tanto representa una recta. Además es la recta en la que se transforma (5) mediante (3). Observemos asimismo que la transformación (3) lleva el eje Ox sobre sí mismo, produciendo sólo un desplazamiento longitudinal de sus puntos. Esto es claro, ya que sobre este eje $y = 0$ y por (3) también $y' = 0$. Sobre el eje Ox la transformación viene dada por una sola fórmula:

$$x' = \frac{x + a}{1 + ax} \quad (|a| < 1). \quad (3')$$

Sobre este eje el segmento x_1x_2 se transforma, por la fórmula (3'), en $x_1x'_2$, y según nuestros convenios sus longitudes se deben considerar iguales. Esta es la forma en que se efectúa el «transporte de segmentos».

Para el centro O del círculo tenemos que $x = 0$ y por tanto $x' = a$; es decir, mediante (3) el centro se traslada a un punto A de coordenada $x = a$.

Puesto que la única condición que se le impone a a es $|a| < 1$, el centro se puede transformar en cualquier punto del diámetro a lo largo del eje Ox .

Mediante la misma transformación el punto que estaba en A se convierte en el punto A' de coordenadas

$$x_1 = \frac{a + a}{1 + a^2} = \frac{2a}{1 + a^2}.$$

Así, el segmento OA se transforma por (3) en AA_1 , y de esta forma es «transportado» sobre la «recta» que representa el diámetro del círculo.

Repetiendo el mismo proceso podemos transportar el mismo segmento un número arbitrario de veces. El punto A_n de coordenada x_n se transforma en el punto A_{n+1} de coordenada

$$x_{n+1} = \frac{x_n + a}{1 + ax_n}.$$

Así obtenemos puntos $A, A_1, A_2, \dots$ de coordenadas

$$x_0 = a, \quad x_1 = \frac{2a}{1 + a^2}, \quad x_2 = \frac{x_1 + a}{1 + ax_1} = \frac{3a + a^3}{1 + 3a^2}, \dots$$

Puesto que todos los segmentos $A_n A_{n+1}$ se obtienen de OA por transformaciones que expresan un movimiento, todos ellos son «iguales» entre sí (iguales en el sentido de la geometría de Lobachevski tal y como está representada en el modelo). Es fácil probar que la sucesión de puntos A_n converge al punto extremo del diámetro. En el sentido del modelo tienden al infinito.

Puesto que el eje Ox se puede tomar en cualquier dirección, las mismas transformaciones de desplazamiento se pueden efectuar a lo largo de cualquier diámetro. Combinándolas con rotaciones alrededor del centro de la circunferencia y con simetrías respecto a un diámetro, obtenemos todos los «movimientos» que se pueden considerar en nuestro modelo. Estas transformaciones se estudiarán con más detalle en la próxima sección, donde probaremos rigurosamente que la geometría de Lobachevski se ve realmente satisfecha por nues-

tro modelo y que, en particular, las transformaciones que hemos definido como movimientos cumplen todas las condiciones (axiomas) a las que están sujetos los movimientos en geometría.

Repitamos una vez más la clase de modelo que propuso Klein para la geometría de Lobachevski. En un plano tomamos el interior de un círculo; un punto se considera como un punto; una recta, como una cuerda (excluyendo los extremos); un movimiento se toma como una transformación que transforma el círculo en sí mismo y las

cuerdas en cuerdas; la situación de los puntos (un punto está sobre una recta; un punto está entre otros dos) se considera con el sentido usual. La regla para medir longitudes y ángulos (y también áreas) se deduce de la forma en que se definen los movimientos; la igualdad de segmentos y ángulos (o de figuras arbitrarias) también se define, y esta misma definición es aplicable a la operación de transportar un segmento a lo largo de otro.

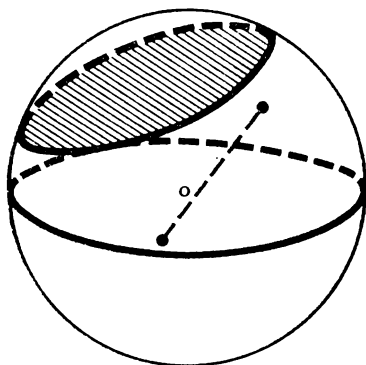


FIG. 13

Con todas estas condiciones, a cada teorema de la geometría de Lobachevski en el plano corresponde un hecho verdadero de la geometría de Euclides dentro del círculo, y viceversa: todo hecho de este tipo se puede reinterpretar en forma de un teorema de la geometría de Lobachevski.

De forma similar se puede construir un modelo de la geometría de Lobachevski en el espacio. Podemos tomar el interior de una esfera (Fig. 13); una recta se interpreta como una cuerda, un plano como un círculo cuya circunferencia esté sobre la esfera; pero la superficie de la esfera, y por tanto los puntos extremos de las cuerdas y las circunferencias de esos círculos, se excluyen; finalmente, un movimiento se define como una transformación de la esfera en sí misma que transforma cuerdas en cuerdas.

Cuando se dio este modelo de la geometría de Lobachevski se estableció al mismo tiempo que su geometría tiene un significado real sencillo. La geometría de Lobachevski es válida porque se puede tomar como una exposición concreta de la geometría en un círculo o en una esfera. Al mismo tiempo se probó su carácter no contradic-

torio: sus resultados no pueden llevar a contradicciones porque cada uno de ellos se puede trasladar al lenguaje de la geometría euclidiana ordinaria dentro del círculo (o una esfera si se trata de la geometría de Lobachevski en el espacio)¹⁰.

3. *Otras interpretaciones*

Después de Klein, Poincaré dio otro modelo de la geometría de Lobachevski, sirviéndose de él para deducir importantes resultados en la teoría de funciones de variable compleja¹¹. En sus manos la geometría de Lobachevski sirvió para resolver difíciles problemas de una rama de las matemáticas completamente distinta. La geometría de Lobachevski ha encontrado además otras aplicaciones en matemáticas y física teórica; por ejemplo, en 1913 el físico Varichak la aplicó a la teoría de la relatividad.

La geometría de Lobachevski se desarrolla con éxito; la teoría de construcciones geométricas, la teoría general de curvas y superficies, la teoría de cuerpos convexos y otras ramas de las matemáticas se desarrollan dentro de este marco.

§5. *Los axiomas de la geometría; su verificación en el presente caso*

1. *Formulación precisa de los axiomas*

Para dar una demostración estrictamente matemática de que el modelo de Klein proporciona realmente una interpretación de la geometría de Lobachevski, hay que determinar exactamente lo que se debe probar. Verificar los teoremas de Lobachevski uno tras otro sería absurdo, pues de hecho hay infinitos, toda vez que siempre se pueden probar otros nuevos. Sin embargo, basta con probar que el modelo de Klein satisface las proposiciones fundamentales de la geometría de Lobachevski, pues de ellas se deducen las restantes. Son estas proposiciones fundamentales las que hay que formular con precisión.

¹⁰ Los matemáticos suelen decir que la geometría de Lobachevski se puede representar dentro de la de Euclides, y por tanto es tan poco contradictoria como ésta.

¹¹ El modelo de Poincaré equivale a lo siguiente: toma también como plano de Lobachevski el interior de un círculo, pero las rectas se interpretan como arcos de un círculo perpendicular a la circunferencia del círculo dado; los movimientos se definen como transformaciones conformes del círculo en sí mismo. (La conexión con las transformaciones conformes da lugar a la relación con la teoría de funciones de una variable compleja.)

Así, el problema de demostrar que la geometría de Lobachevski no es contradictoria se reduce al de formular exacta y completamente sus proposiciones fundamentales, es decir, sus axiomas. Y como los supuestos de la geometría de Lobachevski sólo difieren de los de la de Euclides en el axioma de las paralelas, la tarea consiste en dar una formulación completa y precisa de los axiomas de la geometría de Euclides. Una tal formulación no se encuentra en la obra de Euclides; en particular no existe allí ninguna definición de las propiedades de movimiento o superposición de figuras, aunque, desde luego, hace uso de ellas. El trabajo de construir los axiomas de Euclides de una forma exacta y completa surgió precisamente en conexión con el desarrollo de la geometría de Lobachevski, así como con la ya mencionada tendencia general de finales de siglo a hacer los fundamentos de las matemáticas más rigurosos.

El problema de la formulación de los axiomas de la geometría fue finalmente resuelto gracias a las investigaciones de una serie de geómetras.

Los axiomas se pueden elegir de formas diferentes tomando diferentes conceptos como punto de partida. La relación de los axiomas de la geometría que damos aquí está basada en los conceptos de punto, línea recta, movimiento, y nociones tales como: el punto X está *sobre* la recta a ; el punto B se encuentra *entre* los puntos A y C ; un *movimiento* lleva el punto X sobre el punto Y . (Los demás conceptos se definen en función de éstos; por ejemplo, un segmento se define como el conjunto de puntos que se hallan entre dos puntos dados.)

Los axiomas se dividen en cinco grupos.

I. *Axiomas de incidencia*

1. Por dos puntos pasa una línea recta, y sólo una.
2. Una línea recta contiene al menos dos puntos.
3. Existen al menos tres puntos que no se hallan sobre una recta.

II. *Axiomas de orden*

1. De cada tres puntos que se hallan sobre una línea recta, hay uno que se encuentra entre los otros dos.
2. Si A y B son dos puntos de una línea recta, existe al menos un punto C sobre la recta tal que B se halla entre A y C .
3. Una línea recta divide al plano en dos semiplanos (es decir,

divide aquellos puntos del plano que no están sobre la recta en dos clases tales que los puntos de una misma clase se pueden unir por segmentos que no cortan la recta, mientras que los puntos de distinta clase no).

III. *Axiomas de movimiento*

(Por movimiento debe entenderse no una transformación de una figura concreta, sino de todo el plano.)

1. Un movimiento transforma rectas en rectas.
2. Dos movimientos efectuados sucesivamente son equivalentes a un solo movimiento.
3. Sean A y A' dos puntos, a y a' dos semirrectas que partan de ellos, y α y α' dos semiplanos limitados por las rectas prolongación de a y a' ; existe entonces un único movimiento que lleva A sobre A' , a sobre a' , y α sobre α' . (Hablando intuitivamente, llevamos A sobre A' por medio de traslación, la semirrecta a sobre a' mediante una rotación y finalmente los semiplanos α y α' , o coinciden, o, si no, se efectúa una simetría de eje a .)

IV. *Axioma de continuidad*

1. Sean $X_1, X_2, X_3, \dots$, puntos situados en una línea recta de modo que cada uno de ellos está a la derecha del precedente, y todos ellos a la izquierda de un punto A situado también sobre la recta ¹². Existe entonces un punto B a la derecha de todos los puntos $X_1, X_2, X_3, \dots$, situado también en la recta y tal que al menos un punto X_n está arbitrariamente cerca de él (es decir, cualquiera que sea el punto C a la izquierda de B , existe un punto X_n sobre el segmento CB).

V. *Axioma de paralelismo (Euclides)*

1. Por un punto que no está sobre recta dada sólo se puede trazar una recta que no la corte.

Estos axiomas son suficientes para construir la geometría de Euclides en el plano. Todos los axiomas de un curso escolar se pueden deducir de éstos, aunque la deducción es muy tediosa.

Los axiomas de Lobachevski difieren de los de Euclides solamente en el axioma de paralelismo.

¹² Se puede intercambiar «derecha» por «izquierda».

V'. *Axioma de paralelismo (Lobachevski)*

1. Por un punto que no está en una recta dada pasan al menos dos rectas que no cortan a la primera.

Quizá parezca extraño que en la lista de axiomas se halle, por ejemplo, éste: «una línea recta contiene al menos dos puntos». En nuestra idea de recta va implícito que tiene infinitos puntos. No es extraño, pues, que ni a Euclides ni a ningún otro matemático hasta finales del siglo pasado se les ocurriera que un axioma de este tipo había que formularlo: lo suponían implícitamente. Pero ahora la situación ha cambiado. Al dar una nueva interpretación a la geometría, por recta debemos entender no la línea recta usual, sino cosas bien distintas: una geodésica sobre una superficie, una cuerda de un círculo, o cualquier otra cosa. Surge por tanto la necesidad de formular clara y exhaustivamente *todo* lo que hay que postular acerca de estos objetos que han de ser considerados como rectas. Lo mismo sucede con los demás conceptos y axiomas.

Así, la aparición de las diversas interpretaciones de la geometría, como ya hemos dicho, fue uno de los estímulos más importantes para la formulación exacta de sus principios fundamentales. Este es también el orden histórico: la formulación precisa de los axiomas vino después de los modelos de Beltrami, Klein y Poincaré.

2. *Verificación de los axiomas en el modelo de Klein*

Vamos ahora a demostrar que el modelo de Klein satisface todos los axiomas excepto el axioma de paralelismo de Euclides. Como ya hemos mencionado en la sección precedente (Fig. 11), es evidente que no se cumple este axioma, pero sí el de Lobachevski. Quedan por verificar los axiomas I-IV.

El plano, en el modelo de Klein, es el interior de un círculo (lo tomaremos de radio 1). Los puntos juegan el papel de puntos; las cuerdas, el papel de rectas; las afirmaciones «un punto está sobre una línea recta» y «un punto se encuentra entre otros dos» se deben entender en el sentido usual. Por tanto, es evidente que los axiomas de incidencia, orden y continuidad se satisfacen. Por ejemplo, el tercer axioma de orden significa simplemente que una cuerda divide al círculo en dos partes.

Faltan por verificar los axiomas de movimiento. Un movimiento se define como una transformación del círculo en sí mismo que con-

vierte rectas en rectas. De la definición se deduce que estas transformaciones satisfacen los dos primeros axiomas de movimiento: el primero, porque las líneas rectas son precisamente cuerdas y la conservación de cuerdas significa la de líneas rectas; el segundo, porque si dos transformaciones llevan el círculo sobre sí mismo y cuerdas sobre cuerdas, la transformación resultante también tendrá la misma propiedad; es decir, es uno de los «movimientos» considerados.

Solamente queda entonces por verificar el tercer axioma de movimiento, que es el único punto que implica dificultad.

En primer lugar observemos que el axioma encierra dos afirmaciones.

Sean A y A' dos puntos; a y a' , dos semirrectas que parten de ellos, y α y α' , dos semiplanos limitados por las rectas a y a' .

En primer lugar se exige la existencia de un movimiento que lleve a sobre a' , A sobre A' y α sobre α' .

En segundo lugar, que este movimiento sea único.

Podríamos referirnos al capítulo 3, §14, donde han sido probadas las dos partes del axioma, pero preferimos hacerlo aquí sin involucrar, como en el capítulo 3, otros problemas más generales.

Demostremos que la primera parte es verdadera en nuestro modelo (es decir, con la apropiada interpretación de los términos «semirrecta», «semiplano», «movimiento»).

Para comenzar supongamos que el punto A' está en el centro del círculo. Tomamos un sistema de ejes coordenados con origen en el centro del círculo y el eje Ox pasando por el punto A (Fig. 14). En la sección precedente hemos estudiado la transformación

$$x' = \frac{x + a}{1 + ax}, \quad y' = \frac{y \sqrt{1 - a^2}}{1 + ax}. \quad (6)$$

Allí probamos que era un «movimiento» (es decir, que transforma el círculo en sí mismo y las rectas en rectas).

Sea x_0 la abscisa de A , siendo su ordenada $y_0 = 0$. Por tanto, si tomamos $a = -x_0$, por (6) A se transforma en el punto de coordenadas $(0, 0)$, es decir en A' .

Ahora, puesto que las líneas rectas se transforman en líneas rectas, la «semirrecta» (es decir, el segmento de cuerda) a toma la posición a'' (Fig. 14). Mediante una rotación alrededor del centro podemos llevar a'' sobre a' . El «semiplano» α es uno de los segmentos circulares limitados por la «semirrecta» (cuerda) a . Si después de nuestro movimiento coincide con a' , se ha completado la transfor-

mación; si no, por una simetría de eje el diámetro a' se transforma en a' .

Así, combinando proceso de «desplazamiento» (6) con una rotación y, si es necesario, con una simetría hemos llevado A, a, α sobre A', a', α' . Pero el resultado de todos estos «movimientos» es también un «movimiento» que transforma A, a, α en A', a', α' ; es decir, hemos probado la existencia del movimiento exigido.

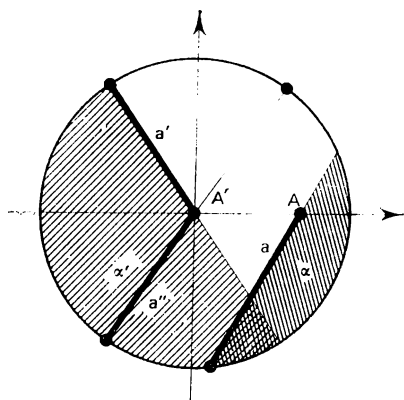


FIG. 14

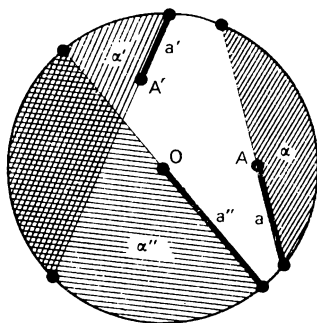


FIG. 15

Hasta ahora nos hemos restringido al caso especial en el que A' estaba en el centro. Supongamos ahora que tiene otra posición cualquiera. Entonces, por lo que ya hemos probado, podemos transformarlo en el centro mediante un «movimiento» que denotamos por D_1 . Entonces la «semirrecta» a' se transforma en otra a'' que pasa por el centro y el «semiplano» α' en un cierto «semiplano» (semicírculo α'') (Fig. 15). Como ya hemos probado, por medio de un cierto movimiento D_2 podemos llevar también A al centro, la «semirrecta» a sobre a'' , el «semiplano» α sobre α'' . Finalmente por el «movimiento» inverso de D_1 llevamos A' a su primitiva posición y al mismo tiempo a'' y α'' vuelven a su posición original a', α' .¹³

Así, como resultado de combinar los «movimientos» D_2 y el

¹³ El «movimiento» inverso de D_1 se expresa mediante las fórmulas (4) (sección 4) si D_1 se expresa por (3).

inverso de D_1 hemos llevado A, a, α sobre A', a', α' . Pero la combinación de dos «movimientos» es otro «movimiento»; y de esta forma hemos comprobado que efectivamente existe un movimiento que lleva A, a, α sobre A', a', α' , cualquiera que sea la posición de A y A' en el círculo. Así, la primera asección contenida en el tercer axioma de movimiento queda completamente probada.

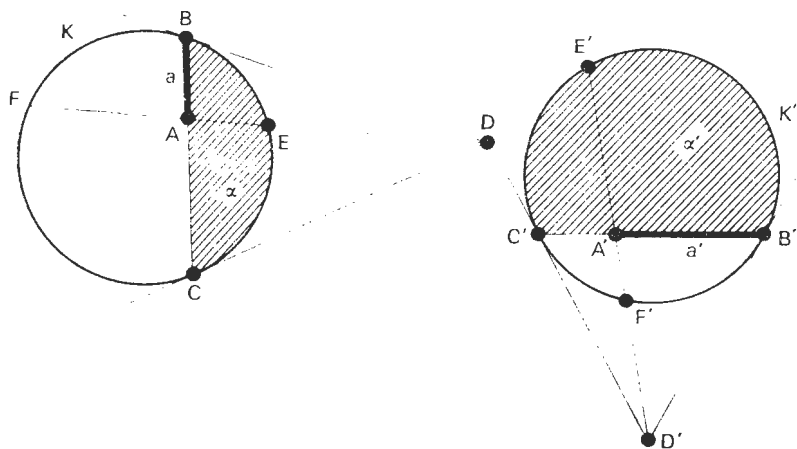


FIG. 16

Mostremos ahora que la segunda asección del axioma también se verifica en nuestro modelo. Para ello debemos probar lo siguiente:

Sean A, A' dos puntos interiores al círculo; a, a' , segmentos de cuerda que arrancan de A y A' , y α, α' , segmentos circulares limitados por las correspondientes cuerdas. Para mayor claridad ilustramos estos puntos, cuerdas y segmentos circulares en dos esquemas diferentes (Fig. 16) aunque, por supuesto, están en el mismo círculo. Se afirma que el «movimiento» que lleva A, a, α sobre A', a', α' , respectivamente, es único, es decir, que está completamente determinado por los datos.

Para la demostración consideremos una transformación no sólo del círculo, sino de todo el plano ¹⁴. Un «movimiento», por definición,

¹⁴ Se puede demostrar que una transformación de un círculo en sí mismo que conserva las líneas rectas se puede extender, de manera única y conservando esta propiedad, a todo el plano proyectivo, es decir, al plano con los puntos del infinito.

transforma rectas en rectas. Una transformación con estas propiedades se llama *proyectiva*. En consecuencia, podemos decir que para nosotros un «movimiento» es una transformación proyectiva que transforma el círculo en sí mismo. (En la figura 16 hemos ilustrado esto haciendo que el círculo K se transforme en el K' . Debemos imaginar solamente que por una traslación el círculo K' se superpone con K .)

Las transformaciones proyectivas se han estudiado en el capítulo 3, sección 13; aquí haremos uso del siguiente teorema probado allí: una transformación proyectiva está completamente determinada por las imágenes de cuatro puntos tales que cada tres de ellos no están alineados.

Volvamos al «movimiento» en cuestión. Lleva el segmento de cuerda a sobre a' y por tanto, B sobre B' y C sobre C' , ya que transforma cuerdas en cuerdas.

Además, puesto que un «movimiento» transforma rectas en rectas y el círculo en sí mismo (en nuestro dibujo, el círculo K en el K'), transforma las tangentes en B y C , en las tangentes en B' y C' . Por tanto, el punto D de intersección de las dos primeras pasa al punto D' de intersección de las segundas¹⁵ (Fig. 16).

Ahora, como A pasa a A' , y las rectas pasan a rectas, la recta AD pasa a $A'D'$. Pero AD corta al círculo en dos puntos E, F , y $A'D'$ lo corta en E', F' . (En el dibujo estos puntos están en las circunferencias K y K' .) Puesto que el círculo se transforma en sí mismo, los puntos E, F , pasan a E', F' . Pongamos E en el arco que limita a , y E' , en el que limita a' . Entonces, como hemos supuesto que a se transforma en a' , E se transforma precisamente en E' y F en F' .

De esta forma hemos encontrado que B, C, E, F se transforman en B', C', E', F' . Ahora bien, es obvio que no puede haber tres puntos de los B, C, E, F y B', C', E', F' , que estén alineados. Por tanto, por el teorema citado anteriormente, la transformación proyectiva que transforma los puntos B, C, E, F en los B', C', E', F' es única. Pero un «movimiento» es una transformación proyectiva, y por consiguiente nuestro «movimiento», que transforma A, a, a en $A' a' a'$, es único, como queríamos demostrar.

Hemos demostrado así que en el modelo en cuestión se satisfacen todos los axiomas de la geometría de Euclides a excepción del de paralelismo, verificándose en cambio el de Lobachevski; o dicho con

¹⁵ Si, por ejemplo, las tangentes en B y C son paralelas, entonces el punto D está «en el infinito».

otras palabras, que todos los axiomas de la geometría de Lobachevski se verifican en el modelo. Este modelo es por tanto una representación de la geometría de Lobachevski. Esta geometría se reduce, por decirlo así, a la geometría de Euclides dentro del círculo, presentada de una manera especial con las interpretaciones convenidas de los términos «línea recta» y «movimiento» en el modelo. Al mismo tiempo esto permite desarrollar la geometría de Lobachevski sobre un modelo concreto; método que resulta más conveniente en muchos problemas.

Desde el punto de vista del análisis lógico de los fundamentos de la geometría, la demostración que hemos dado muestra, primero, que la geometría de Lobachevski no es contradictoria y, segundo, que el postulado de las paralelas no puede ser deducido de los demás axiomas enumerados previamente.

§6. *Separación de las teorías geométricas independientes de la geometría euclidiana*

1. *Geometría proyectiva*

Paralelamente a la creación de la geometría de Lobachevski se llevó a cabo otro desarrollo fundamental más de la geometría. De entre la gran abundancia de propiedades geométricas del espacio, se señalaron y sometieron a estudio independiente distintos grupos de propiedades caracterizados por una estabilidad e interrelaciones peculiares. Estas investigaciones, con sus distintos métodos, dieron lugar a nuevos capítulos de la geometría, es decir, a la ciencia de las formas espaciales, del mismo modo que la anatomía y la fisiología son capítulos distintos en la ciencia del organismo humano.

Inicialmente, la geometría no estaba en absoluto dividida. Estudiaba principalmente las propiedades métricas del espacio relacionadas con la medida de las dimensiones de las figuras. Los problemas no relacionados con la medida, sino con el carácter cualitativo de la posición natural de las figuras, se consideraban solamente de pasada, aunque se venía notando desde hacía tiempo que algunas de estas propiedades se distinguían por una peculiar estabilidad, es decir, se conservaban bajo la acción de deformaciones sustanciales y desplazamientos de situaciones de las figuras.

Consideremos, por ejemplo, la proyección de una figura de un

plano sobre otro (Fig. 17). La longitud de los segmentos y los ángulos varían, y los contornos de los objetos sufren una deformación visible. Sin embargo, se conserva la propiedad de que ciertos puntos están sobre una misma recta, la propiedad de una recta de ser tangente a una curva, etc.

Las proyecciones y las transformaciones proyectivas han sido ya tratadas en el capítulo 3, donde mencionamos su evidente relación

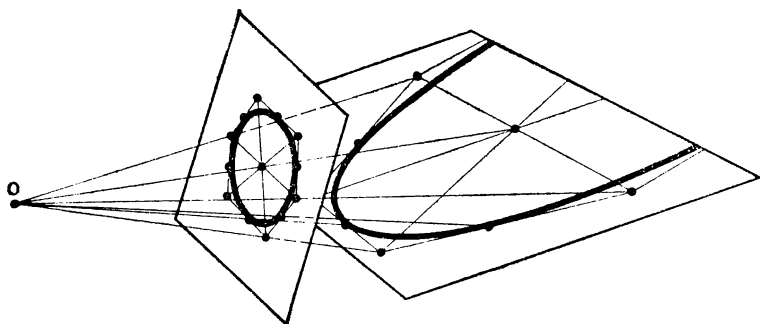


FIG. 17

con la perspectiva, es decir, con el trazado de figuras espaciales sobre un plano. El estudio de las propiedades de la perspectiva se remonta a los tiempos de Euclides y a los trabajos de los antiguos arquitectos; los propios artistas se interesaron por la perspectiva: Durero, Leonardo de Vinci y el ingeniero y matemático Desargues (siglo xvii). Finalmente, a comienzos del siglo xix, Poncelet fue el primero en separar y estudiar sistemáticamente las propiedades que se conservan en una transformación proyectiva arbitraria del plano (o del espacio), creando así una ciencia independiente: la *geometría proyectiva*¹⁶.

Quizá se diga que son muy pocas las propiedades que se conservan en una transformación proyectiva, pero esto no es así en modo alguno. Por ejemplo, no es inmediato que el teorema que dice que los puntos de intersección de los lados opuestos de un exágono

¹⁶ Poncelet, ingeniero militar francés, llevó a cabo sus estudios sobre geometría durante su cautividad en Rusia después de 1812. Su *Traité des propriétés projectives des figures* apareció en 1822.

(prolongados) inscrito en una circunferencia están sobre una recta, sea también válido para una elipse, parábola o hipérbola. El teorema habla solamente de propiedades proyectivas y estas curvas se pueden obtener del círculo por proyección. Menos evidente aún es que el teorema que dice que las diagonales de un exágono circunscrito se cortan en un punto sea consecuencia análoga del teorema; la profunda relación entre ambos solamente se revela mediante la geometría proyectiva. Tampoco resulta evidente que una proyección, que prescinde de la conservación de distancias, conserve constante (para cuatro puntos A, B, C, D) la razón doble $AC/CB : AD/DB$ (Fig. 18):

$$\frac{AC}{CB} : \frac{AD}{DB} = \frac{A'C'}{C'B'} : \frac{A'D'}{D'B'}.$$

Esto significa que en la perspectiva se conservan muchas relaciones. Por ejemplo, haciendo uso de esto, es fácil determinar, mediante una fotografía del camino en que se encuentran, la distancia de los postes de telégrafo A, B, C al punto D (Fig. 18) cuando se conoce la distancia entre dos de ellos.

En el capítulo 3 hemos tratado ya de la geometría proyectiva y de sus aplicaciones a la fotografía aérea. Es razonable que sus leyes se utilicen en arquitectura, en la construcción de panorámicas, en decoración, etc.

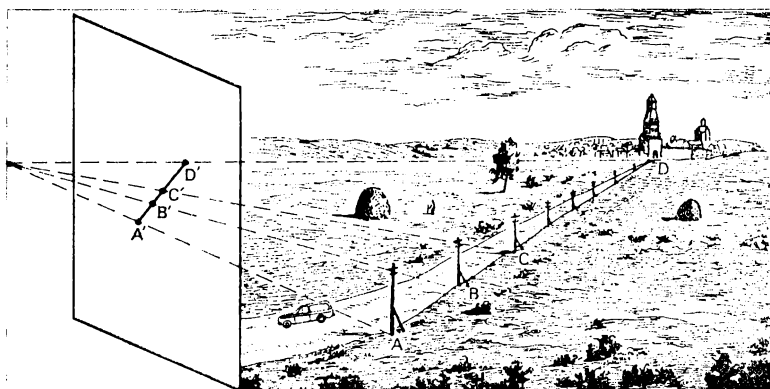


FIG. 18

La emancipación de la geometría descriptiva jugó un importante papel en el desarrollo de la propia geometría.

2. Geometría afín

Podemos tomar como otro ejemplo de geometría independiente la *geometría afín*. Estudia aquellas propiedades de las figuras que permanecen inalteradas en una transformación en la que las coordenadas cartesianas de un punto (x, y, z) y las del transformado (x', y', z') se relacionan por las ecuaciones:

$$\begin{aligned}x' &= a_1x + b_1y + c_1z + d_1, \\y' &= a_2x + b_2y + c_2z + d_2, \\z' &= a_3x + b_3y + c_3z + d_3\end{aligned}$$

(donde suponemos que el determinante

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}$$

es diferente de cero).

Se comprueba que toda transformación afín se reduce a un movimiento, posiblemente a una simetría, en un plano, seguido de una contracción o extensión del espacio en tres direcciones mutuamente perpendiculares.

Estas transformaciones conservan muchas propiedades de las figuras. Las líneas rectas continúan siendo rectas (de hecho se conservan todas las propiedades «proyectivas»); además, las rectas paralelas se transforman en rectas paralelas; se conserva la razón de los volúmenes, así como la razón de las áreas de figuras que están en el mismo plano o en planos paralelos, la razón de las longitudes de segmentos sobre la misma recta o sobre rectas paralelas, etc. Muchos teoremas conocidos pertenecen esencialmente a la geometría afín. Como ejemplos, podemos tomar los teoremas que dicen que las medianas de un triángulo se cortan en un punto, que las diagonales de un paralelogramo se cortan en el punto medio, que los puntos medios de cuerdas paralelas en una elipse se encuentran sobre una línea recta, etc.

Toda la teoría de curvas y superficies de segundo orden está muy relacionada con la geometría afín. La misma división de estas

curvas en elipses, hipérbolas y parábolas se basa, de hecho, en propiedades afines de las figuras: una transformación afín transforma una elipse precisamente en una elipse y nunca en una parábola o hipérbola; análogamente, una parábola se transforma en otra parábola, pero nunca en una elipse, etc.

La importancia del estudio detallado de las propiedades afines generales de las figuras está justificada porque transformaciones incomparablemente más complicadas resultan ser esencialmente lineales, es decir, afines en lo infinitamente pequeño, y la aplicación del cálculo diferencial está precisamente ligada a la consideración de regiones infinitamente pequeñas del espacio.

3. El «Erlanger Programm» de Klein

En 1872 Klein, en una disertación en la Universidad de Erlangen que se conoce como el «Erlanger Programm» resumió los resultados de los estudios sobre geometría proyectiva, geometría afín y otras «geometrías» y dio una clara formulación del principio general de su formación: podemos considerar un grupo¹⁷ arbitrario de transformaciones unívocas del espacio e investigar las propiedades de las figuras que dichas transformaciones conservan.

Desde este punto de vista, las propiedades del espacio están estratificadas respecto a su profundidad y estabilidad. La geometría euclidiana ordinaria hace caso omiso de todas las propiedades de los cuerpos reales, salvo las geométricas; ahora, en las ramas especiales de la geometría, realizamos una abstracción más, al desprestigiar todas las propiedades geométricas excepto las que interesan en una rama concreta de la geometría.

De acuerdo con el principio de Klein, podemos construir muchas geometrías. Por ejemplo, podemos considerar las transformaciones que conservan los ángulos entre líneas arbitrarias (transformaciones conformes del espacio), y al estudiar las propiedades que se conservan en tales transformaciones estamos haciendo geometría conforme. Podemos considerar transformaciones de sólo parte del espacio. Así, al estudiar aquellas transformaciones del círculo en sí mismo que

¹⁷ La palabra «grupo» no se usa aquí en el sentido de una mera colección. Cuando hablamos de un grupo de transformaciones (ver capítulo 19) pensamos en un conjunto de transformaciones que contiene la transformación identidad (deja invariantes los puntos); que contiene, junto con toda transformación no idéntica, la transformación inversa (que coloca los puntos en su primitiva posición), y que, si contiene dos transformaciones, también debe contener a la que resulta de efectuar estas dos una a continuación de la otra.

transforman cuerdas en cuerdas y reunir las propiedades que se conservan en dichas transformaciones, obtenemos una geometría que, como hemos mostrado en los §4 y §5, coincide con la de Lobachevski.

4. *Espacios abstractos*

El ulterior desarrollo de las teorías así separadas, incluso desde el punto de vista teórico (por no hablar de su contenido real), no finalizó en lo que hemos dicho aquí.

Si sólo estamos interesados, por ejemplo, en las propiedades afines de las figuras, podemos imaginar, por abstracción de las demás propiedades, un espacio con figuras que tengan *solamente* las propiedades geométricas que nos interesan, y nada más. En este «espacio» las figuras no tienen más propiedades que las afines. Es natural intentar dar una exposición axiomática de la geometría de tales espacios abstractos, es decir, suponer que tratamos con ciertos objetos abstractos: «puntos», «líneas rectas» y «planos», cuyas propiedades se especifican en ciertos axiomas (hay desde luego menos de esas propiedades que en la geometría euclidiana), de modo que las consecuencias derivadas de esos axiomas correspondan a las propiedades afines de las figuras del espacio ordinario.

Esto es efectivamente factible; una tal colección de «puntos», «líneas rectas» y «planos» abstractos con su sistema de propiedades recibe el nombre de *espacio afín*.

De forma enteramente análoga podemos imaginar un sistema abstracto de objetos que tengan solamente la clase de propiedades correspondientes a las propiedades proyectivas de las figuras en el espacio euclidiano. (Esta vez la diferencia entre el sistema de axiomas y los axiomas de la geometría ordinaria resulta aún más sustancial.)

5. *Espacios topológicos abstractos*

Cuando profundizamos más en la naturaleza de las formas geométricas, constatamos que en un gran número de problemas nos interesan propiedades —incluso más profundas que las proyectivas— tan estrechamente relacionadas con la figura dada que se conservan bajo la acción de deformaciones arbitrarias, a condición solamente de que estas deformaciones no provoquen en la figura roturas ni fusiones. Para conferir a la noción de deformación continua una precisión mayor que la que da la intuición es preciso referirse a la definición de función continua conocida por el análisis. Decimos entonces que se trata de una transformación arbitraria de todos los puntos de una

figura, y que bajo esta transformación las coordenadas cartesianas de los puntos imagen se expresan mediante una función continua de las coordenadas de los puntos originales, al tiempo que las coordenadas antiguas se pueden expresar como funciones continuas de las nuevas.

Las propiedades que se conservan en una transformación arbitraria de este tipo se llaman topológicas y la ciencia que las estudia es la *topología* (ver capítulo 17).

Las propiedades topológicas de una figura (en los casos más simples) se distinguen por su carácter excepcionalmente intuitivo. Por ejemplo, es casi obvio que toda línea del plano obtenida de la circunferencia por deformación continua divide al plano en dos partes, interior y exterior, por mucho que su contorno se haya deformado. Por tanto, la

propiedad de la circunferencia de dividir el plano es topológica. También es, sin duda, obvio que la superficie toroidal (Figura 19) no se puede transformar mediante una deformación continua en una esfera; así, la propiedad

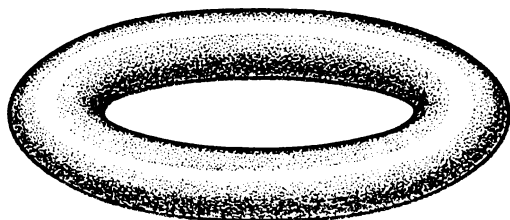


FIG. 19

de una superficie de admitir una deformación continua que la transforme en una superficie toroidal, por ejemplo, es una propiedad topológica que la distingue de otras muchas superficies.

Las discusiones relacionadas con la continuidad son de carácter intuitivo y esclarecen a menudo los conceptos básicos tan bien que resulta muy tentador convertirlas en demostraciones rigurosas y extender tales métodos a otros problemas más complicados.

Tomemos, por ejemplo, los razonamientos que se hacen para probar, en el teorema fundamental del álgebra, el hecho de que toda ecuación

$$z^n + a_1 z^{n-1} + a_2 z^{n-2} + \cdots + a_{n-1} z + a_n = 0 \quad (7)$$

tiene al menos una raíz real o compleja.

Sea z un punto de un plano complejo y $w = f(z)$ el correspondiente punto de otro plano complejo, donde $f(z)$ denota el primer miembro de (7). Para valores absolutos muy grandes de z , la función $f(z)$

difiere relativamente poco de z^n ; pero la función z^n es muy simple. En particular es fácil comprobar que si z , por un movimiento continuo, describe en el plano complejo una circunferencia con centro en el origen, el punto $w_1 = z^n$ da precisamente n vueltas sobre una circunferencia similar de radio $|z|^n$ en el plano w ¹⁸. Entonces el punto $w = f(z)$ describe n vueltas en un contorno Γ' relativamente próximo a la línea descrita por el punto $w_1 = z^n$ (Fig. 20).

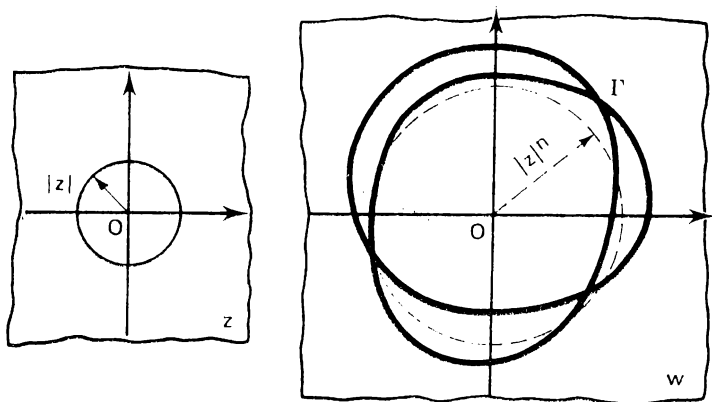


FIG. 20

Cuando la circunferencia descrita por z se contrae continuamente a un punto, el contorno Γ' descrito por $w = f(z)$ en sus n vueltas se deforma continuamente y también se contrae a un punto. Pero es bastante obvio que no puede contraerse a un punto sin pasar por el origen al que inicialmente rodeaba. Por tanto, pasa por el origen al menos una vez y debe existir un z tal que $w = f(z) = 0$. Este z es entonces una raíz de la ecuación (7). Estrictamente hablando está también claro que, en un cierto sentido, debe haber exactamente n raíces, ya que en la contracción cada una de las n vueltas de contorno Γ' pasa por O .

Nuestro razonamiento requiere, por supuesto, una prueba rigurosa de las propiedades topológicas del contorno y sus deformaciones.

¹⁸ Pues si $z = \rho(\cos \phi + i \operatorname{sen} \phi)$ entonces (ver capítulo 4, §3) sabemos que

$$z^n = \rho^n(\cos \phi n + i \operatorname{sen} \phi n);$$

por tanto, cuando el argumento de z varía de 0 a 2π , el de z^n varía de 0 a $2\pi n$ esto es, el radio vector correspondiente a z^n da n vueltas completas.

Podríamos dar muchos ejemplos del uso de propiedades topológicas en diversas ramas de las matemáticas, muy distantes a menudo de la geometría.

Con el estudio de las propiedades topológicas ante nosotros, podemos imaginar un conjunto abstracto de objetos que tengan solamente este tipo de propiedades (ver §7, capítulo 19). Un conjunto de este tipo se llama *espacio topológico abstracto*.

Este punto de vista es ya incomparablemente más amplio que el del estudio de las propiedades topológicas de figuras geométricas solamente. Los espacios topológicos pueden ser muy variados; por ejemplo, los puntos de una superficie toroidal, con sus propiedades específicas de proximidad entre unos y otros, forman un espacio topológico; todos los puntos del plano, otro; el espacio euclidiano completo, un tercero; los puntos de las distintas superficies riemannianas de varias hojas, de las que ya hemos hablado en el capítulo 9, §5, en relación con la teoría de funciones de una variable compleja, forman también otros espacios topológicos. Pero lo más notable es que muchas veces se pueden establecer claramente los conceptos de proximidad y adyacencia (contigüidad) entre objetos que no corresponden en absoluto a nuestra noción de puntos geométricos. Por ejemplo, para todas las posibles posiciones de un mecanismo articulado se puede indicar claramente lo que significa una posición «próxima», o que una posición es «adyacente» a una serie infinita de otras entre las cuales existen posiciones arbitrariamente próximas a una dada.

Vemos, por tanto, que el concepto de espacio topológico es sumamente general. Volveremos sobre este punto en la sección 8.

El objeto de esta sección ha sido dar al lector una idea de las distintas geometrías, mostrar que ciertos problemas concretos conducen al aislamiento e investigación de diversos grupos de propiedades geométricas y que estas investigaciones traen consigo la creación de la idea de objetos geométricos abstractos que tienen *solamente* esas propiedades, es decir, que el aislamiento de esas propiedades en su forma pura conduce a la idea del espacio abstracto correspondiente.

En la siguiente sección se discutirán otros desarrollos que conducen a la construcción de una clase diferente de espacios abstractos.

§7. Espacios multidimensionales

1. La geometría del espacio n -dimensional

Un importante paso en el desarrollo de las nuevas ideas geométricas fue la creación de la geometría de los espacios n -dimensionales a los cuales nos hemos referido ya en el capítulo anterior. Una de las fuerzas motrices fue la tendencia a considerar razonamientos geométricos para la resolución de problemas de álgebra y análisis. El enfoque geométrico a la solución de problemas analíticos se basa en el método de las coordenadas. Demos un sencillo ejemplo.

Queremos saber cuántas soluciones enteras tiene la desigualdad $x^2 + y^2 < N$. Considerando x e y como coordenadas cartesianas en

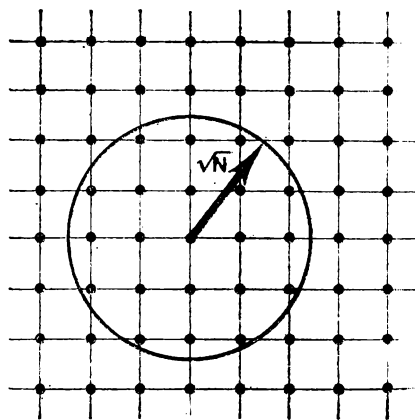


FIG. 21

un plano vemos que el problema se reduce a lo siguiente: ¿cuántos puntos con coordenadas enteras están contenidos en un círculo de radio $\sqrt{N}$? Los puntos de coordenadas enteras son los vértices de los cuadrados de lado unidad que cubren el plano (Fig. 21). El número de tales puntos que están dentro del cuadrado es aproximadamente igual al número de cuadrados contenidos en el círculo, es decir, al área de un círculo de radio $\sqrt{N}$. Así, el número de soluciones de la desigualdad considerada es aproximadamente igual a $\sqrt{N}$. Además, no es

difícil probar que el error relativo cometido tiende a cero cuando $N \rightarrow \infty$. El estudio exacto de este error constituye un problema muy difícil de la teoría de números y ha sido objeto de profundos estudios en años relativamente recientes.

En nuestro ejemplo ha sido suficiente traducir el problema al lenguaje geométrico para obtener en el acto un resultado que no es ni mucho menos obvio desde el punto de vista del «álgebra pura».

El problema correspondiente para una desigualdad con tres incógnitas se resuelve exactamente de la misma forma. Sin embargo, cuando hay más de tres incógnitas el método no es aplicable, ya que nuestro espacio es tridimensional, es decir, la posición de un punto está determinada por una terna de coordenadas. Para conservar la analogía geométrica introducimos entonces la idea de espacio abstracto n -dimensional cuyos puntos están determinados por n coordenadas $x_1, x_2, \dots, x_n$. Los conceptos fundamentales se generalizan de forma que los razonamientos geométricos resultan aplicables a la solución de problemas de n variables; esto facilita mucho la obtención de resultados. La posibilidad de una tal generalización se basa en la uniformidad de las leyes algebraicas, gracias a lo cual se pueden resolver muchos problemas con un número arbitrario de variables. Ello nos permite aplicar razonamientos geométricos que son válidos en tres dimensiones al espacio n -dimensional.

2. Coordenadas en geometría n -dimensional

Los rudimentos del concepto de espacio de cuatro dimensiones aparecen ya en Lagrange, quien, en sus escritos sobre mecánica, consideró formalmente el tiempo como una «cuarta coordenada» junto con las tres coordenadas espaciales. Pero la primera exposición sistemática de una geometría multidimensional fue dada en 1844 por el matemático alemán Grassmann e independientemente por el inglés Cayley, que se apoyaron en una analogía formal con la geometría analítica ordinaria. A continuación damos un esbozo general de esta analogía, en su versión actual.

Un punto en un espacio n -dimensional está determinado por n coordenadas $x_1, x_2, \dots, x_n$. Una figura en un espacio n -dimensional es el lugar geométrico de los puntos que satisfacen ciertas condiciones. Por ejemplo, un «cubo» n -dimensional se define como el lugar geométrico de los puntos cuyas coordenadas cumplen las desigualdades $a \leq x_i \leq b$ ($i = 1, 2, \dots, n$). La analogía con el cubo ordinario es completamente evidente: cuando $n = 3$, es decir, cuando el espacio es tridimensional, las desigualdades definen, de hecho, un cubo cuyas aristas son paralelas a los ejes coordenados y tienen de longitud $b - a$ (la figura 22 ilustra el caso $a = 0, b = 1$).

La distancia de dos puntos se define como la raíz cuadrada de la suma de los cuadrados de las diferencias de coordenadas:

$$d = \sqrt{(x'_1 - x''_1)^2 + (x'_2 - x''_2)^2 + \dots + (x'_n - x''_n)^2}.$$

Esta es una generalización directa de la conocida fórmula de la distancia en el plano o en el espacio, es decir, para $n = 2$ ó 3 .

Ahora es posible definir la igualdad de figuras en el espacio n -dimensional. Dos figuras se consideran iguales si se puede establecer una correspondencia entre sus puntos de forma que la distancia entre pares de puntos correspondientes es siempre la misma. Una transformación que conserva las distancias se dice que es un movimiento generalizado¹⁹. Podemos decir entonces, por analogía con la geometría euclidiana usual, que los

objetos que se estudian en geometría n -dimensional son aquellas propiedades de las figuras que se conservan en los movimientos generalizados. Esta definición del contenido de la geometría n -dimensional quedó establecida en 1870 y proporcionó unos fundamentos precisos para su desarrollo. Desde entonces ha sido objeto de numerosos estudios en direcciones similares a las de la geometría euclidiana (geometría elemental, teoría general de curvas, etc.).

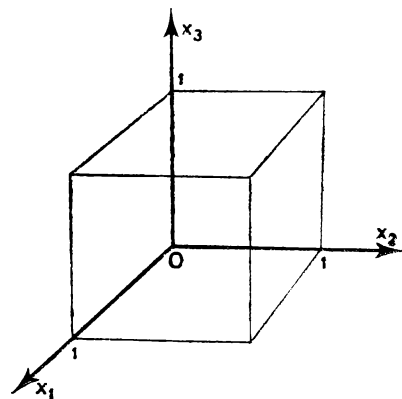


FIG. 22

El concepto de distancia entre dos puntos permite trasladar al espacio n -dimensional otros conceptos de la geometría, tales como segmentos, esfera, longitud, ángulo, volumen, etc. Por ejemplo, una esfera n -dimensional se define como el conjunto de puntos cuya distancia a un punto dado no es mayor que un cierto número R . Por tanto, una esfera se da analíticamente por la desigualdad

$$(x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 \leq R^2$$

donde $a_1, a_2, \dots, a_n$ son las coordenadas del centro. La superficie de la esfera se da por la ecuación

$$(x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 = R^2.$$

¹⁹ La generalización consiste no sólo en el paso a n variables, sino también en el hecho de que las simetrías respecto a un plano se consideran como movimientos, ya que no alteran las distancias entre los puntos.

3. *Hiperplanos*

Por generalización directa damos la siguiente definición: un plano k -dimensional en un espacio n -dimensional es el lugar geométrico de los puntos cuyas coordenadas satisfacen un sistema de $n - k$ ecuaciones lineales.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_1 &= 0, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_2 &= 0 \\ &\vdots \\ a_{n-k,1}x_1 + a_{n-k,2}x_2 + \cdots + a_{n-k,n}x_n + b_{n-k} &= 0. \end{aligned} \quad (8)$$

supuesto que las ecuaciones son compatibles e independientes (es decir, ninguna de ellas es consecuencia de las demás). Cada una de estas ecuaciones representa un hiperplano $(n - 1)$ -dimensional, y en conjunto determinan los puntos comunes a los $n - k$ hiperplanos. Que las ecuaciones (8) sean compatibles significa que existen puntos que las satisfacen, es decir, que los $n - k$ hiperplanos $(n - 1)$ -dimensionales dados se cortan. Que ninguna de las ecuaciones sea consecuencia de las demás significa que ninguna de ellas se puede omitir. De otro modo el sistema se podría reducir a otro de menor número de ecuaciones y definiría un hiperplano de mayor dimensión. Así, pues, geoméricamente un hiperplano k -dimensional es la intersección de $n - k$ hiperplanos $(n - 1)$ -dimensionales representados por ecuaciones independientes. En particular, cuando $k = 1$, tenemos $n - 1$ ecuaciones que determinan un «hiperplano unidimensional», es decir, una recta. Así, esta definición de hiperplano es una generalización formal muy natural de los resultados de la geometría analítica. La ventaja

2. Si dos hiperplanos de dimensiones l y k respectivamente en un espacio n -dimensional tienen al menos un punto en común y $l + k \geq n$, entonces se cortan en un hiperplano de dimensión no menor que $l + k - n$.

De aquí se sigue, como caso especial, que, si dos planos bidimensionales en un espacio tridimensional no coinciden y no son paralelos, entonces se cortan según una recta ($n = 3$, $l = 2$, $l + k - n = 1$). Pero en un espacio de cuatro dimensiones dos planos bidimensionales pueden tener un único punto en común. Por ejemplo, los planos dados por los sistemas de ecuaciones:

$$\left. \begin{array}{l} x_1 = 0 \\ x_2 = 0 \end{array} \right\}, \quad \left. \begin{array}{l} x_3 = 0 \\ x_4 = 0 \end{array} \right\},$$

es evidente que sólo se cortan en el punto de coordenadas $x_1 = 0$, $x_2 = 0$, $x_3 = 0$, $x_4 = 0$.

La demostración del teorema es extremadamente simple: un hiperplano l -dimensional viene dado por $n - l$ ecuaciones, y uno k -dimensional, por $n - k$; las coordenadas de los puntos de intersección deben satisfacer simultáneamente estas $(n - l) + (n - k) = n - (l + k - n)$ ecuaciones. Si ninguna de las ecuaciones es consecuencia de las demás, por la misma definición de hiperplano obtenemos como intersección un hiperplano $(l + k - n)$ -dimensional; de otra forma tendremos un hiperplano de dimensión mayor.

A estos dos teoremas podemos añadir otros dos.

3. En cada hiperplano k -dimensional existen al menos $k + 1$ puntos que no están en un mismo hiperplano de menor dimensión. En un espacio n -dimensional existen al menos $n + 1$ puntos que no pertenecen a ningún hiperplano.

4. Si una recta tiene dos puntos en común con un hiperplano (de dimensión arbitraria), entonces está contenida en el hiperplano. De modo general, si un hiperplano l -dimensional tiene $l + 1$ puntos en común con un hiperplano k -dimensional y esos $l + 1$ puntos no están contenidos en un hiperplano $(l - 1)$ -dimensional, entonces el primer hiperplano está contenido en dicho hiperplano k -dimensional.

Hagamos notar que la geometría n -dimensional se puede construir partiendo de unos axiomas que generalicen la lista del §5. Con este enfoque, los cuatro teoremas mencionados aquí asumen el papel de axiomas de incidencia. Esto muestra, de paso, que el concepto de axioma es relativo: la misma aserción puede ser un teorema en la construcción de una teoría y un axioma en otra.

4. *Ejemplos de espacio n -dimensional*

Hasta aquí hemos dado una idea general del concepto matemático de espacio n -dimensional. A fin de aclarar el significado físico de este concepto, volvamos otra vez al problema de la representación gráfica. Supongamos, por ejemplo, que deseamos representar la dependencia entre la presión de un gas y su volumen. Tomamos un sistema de ejes coordenados en el plano, representando el volumen v en uno de los ejes y la presión p en el otro. La dependencia entre la presión y el volumen en unas condiciones dadas está representada por una cierta curva (por la conocida ley de Boyle-Mariotte esta curva sería, a temperatura fija, una hipérbola). Pero cuando tenemos un sistema físico más complicado, cuyo estado viene dado no ya por dos datos (como presión y volumen en el caso de un gas), sino, por ejemplo, mediante cinco, entonces la representación gráfica de su conducta conduce a la noción de espacio de cinco dimensiones.

Supongamos, por ejemplo, que estamos interesados en una aleación de tres metales o una mezcla de tres gases. El estado de la mezcla se determina mediante cuatro datos: la temperatura T , la presión p , y los contenidos porcentuales c_1, c_2 de dos gases (el porcentaje del tercero está determinado por el hecho de que la suma total de los porcentajes es 100 por 100, de modo que $c_3 = 100 - c_1 - c_2$). El estado de una mezcla así viene determinada, pues, por cuatro datos. Una ilustración gráfica del sistema requiere, o bien una combinación de varios diagramas, o bien una representación del estado de la mezcla por medio de un punto de un espacio de cuatro dimensiones de coordenadas T, p, c_1, c_2 . Representaciones de este tipo son efectivamente usadas en química; la aplicación de los métodos de la geometría multidimensional a la química fue desarrollada por el científico americano Gibbs y por la escuela de fisicoquímicos soviéticos del académico Kurnakov. Aquí, la introducción de un espacio multidimensional viene impuesta por el deseo de conservar las provechosas analogías y útiles razonamientos geométricos que se derivan del simple artificio de la representación gráfica.

Demos ahora un ejemplo en el campo de la geometría. Una esfera está determinada por cuatro datos: las tres coordenadas de su centro y su radio ²¹. La geometría especial de esferas, que fue construida

²¹ Nota del traductor al inglés: por consiguiente una esfera se puede representar como un punto de un espacio tetradimensional.

aproximadamente hace un siglo, se puede considerar por tanto como una geometría tetradimensional.

De lo que llevamos dicho se deduce que la base real de la introducción del concepto de espacio multidimensional fue muy clara. Si una figura, el estado de un sistema, etc., está determinado por n datos, entonces esa figura o ese estado, etc., se puede considerar como un punto de un espacio n -dimensional. La ventaja de esta representación viene a ser la misma que la de los gráficos ordinarios: consiste en la posibilidad de aplicar los métodos y analogías de la geometría al estudio del fenómeno en cuestión.

No existe por tanto ningún misticismo en el concepto matemático de espacio multidimensional. No es más que un cierto concepto abstracto creado por los matemáticos con el propósito de describir en lenguaje geométrico aquellas cosas que no admiten una ilustración geométrica sencilla en el sentido usual. Este concepto abstracto tiene una base enteramente real, refleja la realidad y fue creado por las demandas de la ciencia y no por un sueño ocioso de la imaginación. Refleja el hecho de que existen cosas tales como una esfera o una mezcla de gases que vienen caracterizados por una serie de datos que hacen que la colección de todas esas cosas sea multidimensional. El número de variables en un caso dado es precisamente el número de datos; al igual que un punto modifica sus tres coordenadas al moverse en el espacio, una esfera que se mueva, se expanda y se contraiga modifica sus cuatro «coordenadas», es decir, las cuatro cantidades que la determinan.

En las siguientes secciones insistiremos sobre la geometría multidimensional. Es importante comprender aquí que se trata de un método de descripción matemática de cosas y fenómenos reales. La idea de que existe un espacio de cuatro dimensiones en el que está sumergido nuestro espacio real ha sido creada por ciertos literatos y espiritualistas y no tiene ninguna relación con el concepto matemático de espacio tetradimensional. De hablar aquí de alguna relación con la ciencia, ello sólo es posible en el sentido de una deformación imaginaria de los conceptos científicos.

5. *Poliedros en el espacio n -dimensional*

Como ya hemos dicho, la geometría del espacio multidimensional fue construida, en principio, a través de una generalización formal de la geometría analítica usual a un número arbitrario de variables. Sin embargo, tal enfoque no podía satisfacer por completo a los matemáticos. La realidad es que el propósito fue, no tanto una generaliza-

ción de los conceptos geométricos, como de los métodos geométricos de investigación mismos. Por tanto, era importante realizar una exposición puramente geométrica de la geometría n -dimensional independientemente del aparato analítico. Esto fue hecho por primera vez por el matemático suizo Schläefli en 1852, uno de cuyos artículos trata sobre poliedros regulares en un espacio multidimensional. Ciertamente, el artículo de Schläefli no fue valorado justamente por sus contemporáneos, ya que para comprenderlo era necesario elevarse, en cierta medida, a un punto de vista geométrico muy abstracto. Sólo posteriores desarrollos de la matemática han arrojado completa claridad sobre este problema mediante una elucidación exhaustiva de las relaciones mutuas entre los puntos de vista analítico y geométrico. Puesto que no podemos profundizar en este problema, nos limitaremos a algunos ejemplos de la exposición geométrica de la geometría n -dimensional.

Consideremos la definición geométrica de un cubo n -dimensional. Cuando movemos un segmento en un plano a lo largo de una dirección perpendicular a él una longitud igual a la suya engendramos un cuadrado, es decir, un cubo bidimensional (Fig. 23 a). Análogamente, cuando desplazamos un cuadrado en la dirección perpendicular a su propio plano una longitud igual a su lado engendramos un cubo tridimensional (Fig. 23 b). Para obtener un cubo de cuatro dimensiones operamos de la misma forma: tomamos un hiperplano tridimensional en un espacio de cuatro dimensiones; en él, un cubo tri-

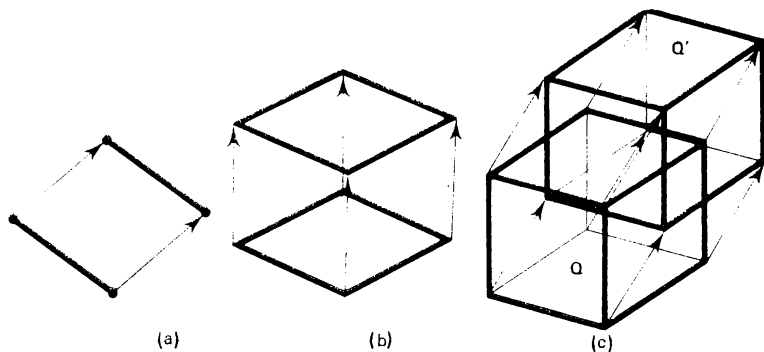


FIG. 23

dimensional, y lo desplazamos en una dirección perpendicular a su hiperplano una distancia igual a la longitud de sus lados (por definición, una línea recta es perpendicular a un hiperplano k -dimensional si es perpendicular a todas las rectas de ese hiperplano). Esta construcción está simbólicamente representada en la figura 23 c, donde se han dibujado dos cubos Q y Q' , que son, respectivamente, las posiciones inicial y final del cubo dado. Los segmentos que unen los vértices representan los segmentos trazados por los vértices del cubo en su movimiento. Vemos que un cubo tetradimensional tiene en total 16 vértices: 8 para el cubo Q y 8 para Q' . Tiene además 32 aristas: 12 en el cubo que se desplaza, en su posición inicial Q , 12 en la posición final Q' y 8 aristas «laterales». Tiene 8 caras tridimensionales, que son a su vez cubos. En el desplazamiento del cubo tridimensional cada una de sus caras engendra otro cubo tridimensional, obteniendo así 6 cubos como caras laterales del cubo tetradimensional; por otra parte hay dos caras más: la «frontal» y la «posterior», que corresponden a las posiciones inicial y final del cubo dado. Finalmente, nuestro cubo tetradimensional tiene 24 caras bidimensionales: 6 en cada uno de los cubos Q y Q' , y otras 12 engendradas por las aristas de Q en su movimiento.

Así, pues, un cubo tetradimensional tiene 8 caras tridimensionales, 24 bidimensionales, 32 unidimensionales (aristas) y finalmente 16 vértices; cada cara es un «cubo» de la dimensión apropiada: un cubo tridimensional, un cuadrado, un segmento y un vértice (que podemos considerar como un cubo de dimensión cero).

De la misma forma, desplazando un cubo tetradimensional «en la quinta dimensión» obtenemos el cubo de dimensión cinco, y así, repitiendo la construcción, podemos construir un cubo de dimensión arbitraria. Todas las caras de un cubo n -dimensional son a su vez cubos de dimensión más pequeña: $(n - 1)$ -dimensionales, $(n - 2)$ -dimensionales, etc., y finalmente unidimensionales, es decir, aristas. Para el lector curioso no será difícil hallar el número de caras de cada dimensión en un cubo n -dimensional. Es fácil comprobar que el número de caras $(n - 1)$ -dimensionales es $2n$ y el de vértices 2^n . ¿Pero cuántas aristas tiene, por ejemplo?

Consideremos otro poliedro en el espacio n -dimensional. En el plano el polígono más simple es el triángulo; tiene el menor número posible de vértices. Para obtener un poliedro con el menor número de vértices basta tomar un punto en un plano distinto del del triángulo y unir dicho punto mediante segmentos con cada uno de los puntos del triángulo. Los segmentos obtenidos llenan una pirámide de tres

lados, es decir, un tetraedro (Fig. 24). Para obtener el poliedro más sencillo en un espacio de dimensión cuatro, razonamos como sigue. Tomamos en un hiperplano tridimensional un cierto tetraedro T . Ahora tomamos un punto fuera del hiperplano y unimos, mediante segmentos, el punto con todos y cada uno de los puntos del tetraedro T . En la figura 24, a la derecha, hemos ilustrado simbólicamente esta construcción. Cada uno de los segmentos que unen el punto O con uno de los puntos del tetraedro no tiene ningún otro punto en común con éste, ya que de otro modo estaría totalmente contenido en el hiperplano tridimensional que contiene a T . Todos estos segmentos,

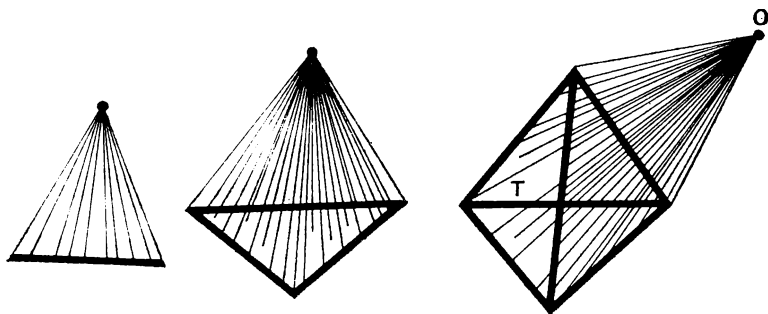


FIG. 24

por decirlo así, están «en la cuarta dimensión» y forman el poliedro tetradimensional más simple llamado *simplex tetradimensional*. Sus caras tridimensionales son tetraedros: uno en la base y 4 caras laterales apoyándose en las caras bidimensionales de la base; en conjunto 5 caras. Sus caras bidimensionales son triángulos; en total hay 10: 4 en la base y 6 laterales. Por último, tiene 10 aristas y 5 vértices.

Repitiendo la misma construcción para un número cualquiera n de dimensiones obtenemos el poliedro n -dimensional más sencillo, que recibe el nombre de *simplex n -dimensional*. Como es evidente por la construcción, tiene $n + 1$ vértices. Se ve fácilmente que todas sus caras son también simples de dimensión más pequeña: $(n - 1)$ -dimensionales, $(n - 2)$ -dimensionales, etc.²²

También son fáciles de generalizar los conceptos de prisma y pirá-

²² Cada m vértices de un simplex determinan el simplex $(m - 1)$ -dimensional generado por ellos: una cara $(m - 1)$ -dimensional del simplex n -dimensional dado. El número de caras $(m - 1)$ -dimensionales de un simplex n -dimensional es por

mide. Si desplazamos paralelamente un polígono plano obtenemos un prisma. Análogamente, desplazando un poliedro tridimensional en la cuarta dimensión engendramos un prisma tetradimensional (ilustrado simbólicamente en la Fig. 25). Un cubo tetradimensional es, evidentemente, un caso particular de prisma.

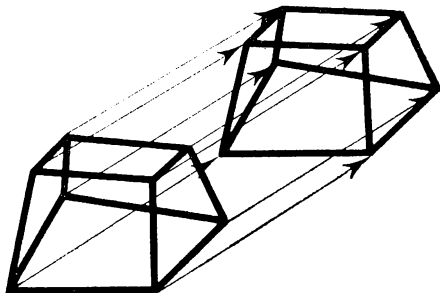


FIG. 25

Una pirámide se construye como sigue. Tomamos un polígono Q y un punto O que no esté en su plano. Unimos cada punto de Q mediante un segmento a O ; estos segmentos llenan una pirámide de base Q (Fig. 26).

Del mismo modo, dados un poliedro Q tridimensional en un espacio tetradimensional, y un punto O que no pertenezca al hiperplano de Q , los segmentos que unen O con los puntos de Q forman una pirámide tetradimensional de base Q . Un simplex tetradimensional no es más que una pirámide que tiene como base un tetraedro.

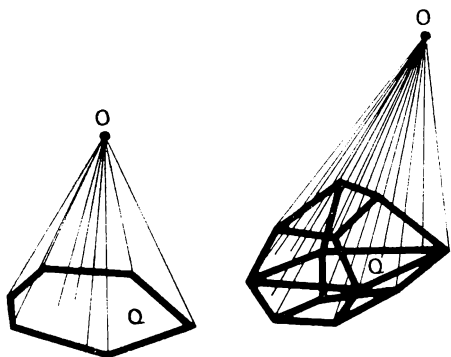


FIG. 26

Exactamente de la misma forma, partiendo de un poliedro Q $(n - 1)$ -dimensional podemos definir un prisma y una pirámide n -dimensionales.

En general, un poliedro n -dimensional es una parte de un espacio n -dimensional

limitada por un número finito de porciones de hiperplanos $(n - 1)$ -dimensionales; un poliedro k -dimensional es una parte de un

tanto igual al número de combinaciones de $n + 1$ de sus vértices tomados de m en m , es decir,

$$C_m^{n+1} = \frac{(n + 1)!}{m!(n - m + 1)!}$$

hiperplano k -dimensional limitada por un número finito de partes de hiperplanos $(k - 1)$ -dimensionales. Las caras de un poliedro son a su vez poliedros de un número menor de dimensiones.

La teoría de poliedros n -dimensionales es una generalización de la de poliedros tridimensionales ordinarios y está llena de resultados concretos. En algunos casos los teoremas sobre poliedros tridimensionales se generalizan sin mayores dificultades a un número arbitrario de dimensiones, pero también existen problemas cuya solución para poliedros n -dimensionales presenta enormes dificultades. Debemos mencionar aquí las profundas investigaciones de C. F. Voronoi (1868-1908), que surgieron, dicho sea de paso, en relación con problemas de la teoría de números; sus trabajos fueron continuados por los geómetras soviéticos. Uno de los problemas, llamado «problema de Voronoi», no está aún completamente resuelto²³.

Podemos tomar los poliedros regulares como ejemplo de las diferencias esenciales que existen entre espacios de diferentes dimensiones. En el plano, un poliedro regular puede tener cualquier número de lados. En otros términos, existe un número definido de «poliedros bidimensionales» distintos. Por otra parte, existen en total cinco

formas de poliedros regulares tridimensionales: el tetraedro, el cubo, el octaedro, el dodecaedro y el icosaedro. En el espacio tetradimensional existen seis formas de poliedros regulares, pero en cualquier espacio de dimensión mayor hay solamente tres, que son: 1) el análogo al tetraedro, el simplex n -dimensional regular, es decir, el simplex cuyas aristas son todas iguales; 2) el cubo n -dimensional; 3) el análogo al octaedro, que se construye como sigue: los centros de las caras del cubo son los vértices del nuevo poliedro.

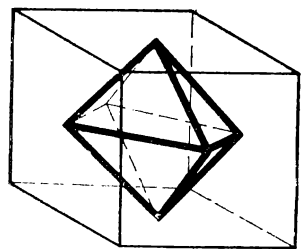


FIG. 27

Para el caso del espacio tridimensional esta construcción se ilustra en la figura 27. Vemos así que en lo que concierne a poliedros re-

²³ Se refiere a la determinación de aquellos poliedros convexos que pueden llenar el espacio uniéndolos entre sí paralelamente cara con cara. En el caso del espacio tridimensional el problema fue planteado y resuelto por Fedorov en relación con las necesidades de la cristalografía; Voronoi y sus sucesores hicieron algunos progresos en el mismo problema para el espacio n -dimensional, pero la solución final se conoce solamente para espacios de dos, tres y cuatro dimensiones.

gulares los espacios de dos, tres y cuatro dimensiones ocupan un lugar especial.

6. Cálculo de volúmenes

Discutamos ahora el problema del volumen de un cuerpo en un espacio n -dimensional. El volumen de un cuerpo n -dimensional se define de forma similar a como se hace en la geometría ordinaria. Es una característica numérica ligada a la figura y se postula que cuerpos iguales tienen volúmenes iguales (es decir, que el volumen no cambia cuando la figura se mueve como un sólido rígido), y que, si un cuerpo está formado por otros varios, su volumen es igual a la suma de los volúmenes de las partes. Como unidad de volumen tomamos el volumen del cubo con aristas de longitud la unidad. Se prueba entonces que un cubo de aristas a tiene un volumen a^n . La demostración se hace exactamente igual que en el plano y en el espacio tridimensional llenando el cubo con capas de cubos más pequeños (Fig. 28). Puesto que los cubos se colocan en n direcciones, obtenemos a^n .

Para definir el volumen de un cuerpo n -dimensional arbitrario, se le sustituye por una figura aproximada compuesta de numerosos cubos n -dimensionales, de forma similar a como en la figura 29 se ha sustituido una figura plana por una figura de cuadrados. El volumen de un cuerpo se define como el límite del volumen de los cuerpos aproximados cuando el tamaño de los cubos que los integran decrece infinitamente.

El volumen k -dimensional de una figura k -dimensional contenida en cierto hiperplano k -dimensional se define exactamente de la misma forma. De la definición de volumen es fácil deducir una importante

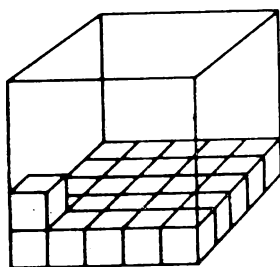


FIG. 28

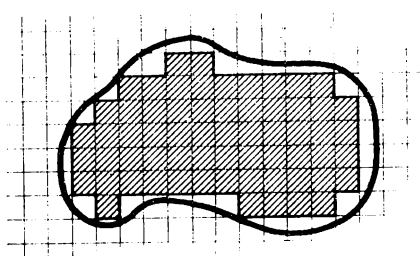


FIG. 29

propiedad: si un cuerpo crece proporcionalmente, es decir, si todas sus dimensiones lineales se incrementan λ veces, su volumen k -dimensional se incrementa λ^k veces.

Si se divide un cuerpo en capas paralelas, entonces su volumen es la suma de los volúmenes de dichas capas:

$$V = \sum V_i.$$

El volumen de cada capa se puede representar aproximadamente como el producto de su altura Δh_i por el volumen $(n - 1)$ -dimensional («área») de la correspondiente sección S_i . Como resultado obtenemos que el volumen total del cuerpo se puede representar aproximadamente por la suma

$$V \approx \sum S_i \Delta h_i.$$

Pasando al límite cuando $\Delta h_i \rightarrow 0$, obtenemos la expresión del volumen en forma de integral

$$V = \int_0^H S(h) dh, \quad (9)$$

donde H es la longitud del cuerpo en la dirección perpendicular a la sección.

Todo esto es completamente análogo al cálculo de volúmenes en el espacio tridimensional. Por ejemplo, en un prisma todas las secciones son iguales y su volumen por tanto no depende de h . De aquí que para el prisma $V = SH$; es decir, el volumen de un prisma es igual al producto del «área» de la base por la altura. Determinemos el volumen de una pirámide n -dimensional. Suponemos que el área de la base es S , y H , la altura. La cortamos por un plano paralelo a la base a una distancia h del vértice. De este modo hemos separado una pirámide de altura h . Denotemos el área de la base por $s(h)$. Esta pirámide más pequeña es, evidentemente, semejante a la original: todas sus dimensiones están disminuidas en la razón de h a H , es decir, están multiplicadas por h/H . Por tanto, el volumen $(n - 1)$ -dimensional («área») de su base es

$$s(h) = \left(\frac{h}{H} \right)^{n-1} S,$$

ya que cuando las dimensiones lineales de una figura $(n - 1)$ -dimensional se multiplican por un factor λ , el volumen queda multiplicado por λ^{n-1} .

Por la fórmula (9) el volumen de toda la pirámide es igual a

$$V = \int_0^H s(h) dh,$$

de donde

$$V = \int_0^H S \left(\frac{h}{H} \right)^{n-1} dh = \frac{S}{H^{n-1}} \int_0^H h^{n-1} dh = \frac{S}{H^{n-1}} \cdot \frac{H^n}{n} = \frac{1}{n} SH;$$

es decir, el volumen de una pirámide n -dimensional es igual a la n -ésima parte del producto del «área» [volumen $(n - 1)$ -dimensional] de la base por la altura. Para $n = 2$ ó 3 obtenemos, como casos especiales, resultados conocidos: el área (volumen bidimensional) de un triángulo es igual a la mitad del producto de la base por la altura, y el volumen de una pirámide tridimensional es un tercio del producto del área de la base por la altura.

En cuanto a la esfera, podemos suponer que está aproximadamente compuesta por un sinfín de estrechas pirámides que tienen como vértice común el centro de la esfera. La altura de dichas pirámides es igual al radio de la esfera y la superficie de sus bases σ_i cubre toda la superficie de la esfera aproximadamente. Puesto que el volumen de cada pirámide es $\frac{1}{n} R \sigma_i$, sumando todos los volúmenes obtenemos que el volumen de la esfera es

$$V \approx \frac{1}{n} R \sum \sigma_i \approx \frac{1}{n} RS.$$

En el límite obtenemos la fórmula exacta: $V = \frac{1}{n} RS$; es decir, el volumen de la esfera es igual a la n -ésima parte del producto de su radio por su superficie. Para n igual a dos o tres la relación es muy conocida ²⁴.

Mencionemos una propiedad de la esfera que se puede probar para el espacio n -dimensional exactamente de la misma forma que para el espacio tridimensional: entre todos los cuerpos de un volumen dado la esfera es la de menor superficie.

²⁴ El cálculo del volumen de una esfera se puede efectuar también aplicando la fórmula (9); una sección de una esfera n -dimensional es una esfera $(n - 1)$ -dimensional, y por consiguiente el volumen de una esfera n -dimensional se puede calcular pasando de $(n - 1)$ a n .

7. Geometría n -dimensional «superior»

Hasta ahora nos hemos limitado a la geometría elemental del espacio n -dimensional; pero podemos desarrollar también la geometría «superior», por ejemplo, la teoría general de curvas y superficies. En un espacio n -dimensional las superficies pueden ser de distintas dimensiones: «superficies» unidimensionales, es decir, curvas; superficies bidimensionales, tridimensionales, . . . , y por último $(n - 1)$ -dimensionales. Una curva se puede definir como el lugar geométrico de los puntos cuyas coordenadas dependen continuamente de un parámetro variable t :

$$x_1 = x_1(t), \quad x_2 = x_2(t), \dots, \quad x_n = x_n(t).$$

Una curva es la trayectoria, por decirlo así, del movimiento de un punto en un espacio n -dimensional al variar t . Si nuestro espacio sirve como ilustración del estado de algún sistema físico, tal como los que hemos discutido en la subsección 4, entonces una curva representa una sucesión continua de estados, o el curso del cambio de estado en relación con el parámetro t (por ejemplo, el tiempo). Esto generaliza la representación gráfica usual del proceso de cambio de estado por medio de curvas.

A cada punto de una curva en un espacio n -dimensional asociamos no sólo una tangente («plano tangente unidimensional»), sino también hiperplanos tangentes de todas las dimensiones desde 2 hasta $n - 1$. La velocidad de rotación de cada uno de estos hiperplanos, respecto a la velocidad de incremento de longitud del arco, da la correspondiente curvatura. Así, una curva tiene $n - 1$ hiperplanos tangentes, desde el unidimensional hasta el $(n - 1)$ -dimensional, y posee $n - 1$ curvaturas. La geometría diferencial en un espacio n -dimensional resulta mucho más complicada que en el espacio tridimensional.

Hasta ahora sólo hemos hablado de la geometría n -dimensional que es generalización directa de la geometría euclidiana ordinaria. Pero ya sabemos que, aparte de esta geometría, existen también la geometría de Lobachevski, la proyectiva y otras. Estas geometrías se generalizan fácilmente, del mismo modo, a un número arbitrario de dimensiones.

§8. *Generalización del alcance de la geometría*1. *El espacio de los colores*

Cuando en la sección anterior hablábamos del significado real del espacio n -dimensional, topamos con el problema de la generalización del alcance de la geometría, con el problema del concepto general de espacio en matemáticas. Pero antes de dar la correspondiente definición general, consideremos algunos ejemplos.

La experiencia muestra —y esto fue ya indicado por M. V. Lomonosov— que la visión humana normal es tricromática, es decir, toda percepción cromática de un color C es una combinación de tres colores fundamentales: rojo R , verde V y azul A , con intensidades específicas²⁵. Si denotamos estas intensidades, en ciertas unidades, por x , y , z , podemos escribir $C = xR + yV + zA$. Igual que un punto en el espacio se puede mover de arriba a abajo, de izquierda a derecha y de delante hacia atrás, la percepción de color, de un color C , puede variar continuamente al variar en las tres direcciones sus partes componentes, rojo, verde y azul. En consecuencia, podemos decir, por analogía, que el conjunto de todos los colores posibles es «el espacio de color tridimensional». Las intensidades x , y , z juegan el papel de coordenadas de un punto, de un color C . (Una diferencia importante respecto a las coordenadas ordinarias consiste en que las intensidades no pueden ser negativas. Cuando $x = y = z = 0$ obtenemos un color perfectamente negro correspondiente a una ausencia completa de luz.)

Un cambio continuo de color se puede representar por una línea en el «espacio de color»; los colores del arco iris, por ejemplo, forman una línea de este tipo; también forman una línea de color las percepciones que se obtienen de un objeto de coloración homogénea sometido a un cambio continuo en el brillo de la iluminación. En este caso sólo cambia la intensidad de la percepción; su coloración permanece invariable.

²⁵ Se trata aquí de la percepción de color, no de luz. La percepción de color es un fenómeno objetivo—una reacción a la luz. Una misma percepción se puede producir mediante luz de distintas longitudes de onda. Por ejemplo, el color verde se puede obtener no sólo a partir de luz verde espectralmente pura, sino también como mezcla de rojo y azul. Por otra parte, las personas que padecen de daltonismo tienen solamente dos percepciones fundamentales; los casos de daltonismo en que sólo se da una sola percepción fundamental de color son muy raros.

Además, dados dos colores, pongamos rojo R y blanco B , si los mezclamos en proporciones variables ²⁶, obtenemos una gama de colores continua, desde el rojo R hasta el blanco B , que podemos llamar el segmento RB . La idea de que el color rosa está entre el rojo y el blanco tiene así un significado claro.

De este modo surge el concepto de las figuras y relaciones geométricas más sencillas en el «espacio de color». Un «punto» es un color, el «segmento» AB es el conjunto obtenido por la mezcla de los colores A y B ; la aserción «el punto D está en el segmento AB » significa que D es una mezcla de A y B . La mezcla de tres colores da un trozo de plano —un «triángulo de color». Todo esto se puede describir analíticamente usando las coordenadas de color x, y, z , y las fórmulas que dan las líneas de color y los planos de color son enteramente análogas a las de la geometría analítica ordinaria ²⁷.

En el espacio de color, las relaciones de la geometría de Euclides en lo que toca a la disposición de puntos y segmentos se satisfacen. El sistema de estas relaciones forma una geometría afín, y se puede decir que el conjunto de todas las posibles percepciones de color representan un espacio afín. (Esto no es completamente exacto, ya que, como hemos dicho, las coordenadas de color x, y, z no pueden ser negativas. Por tanto, el espacio de color corresponde solamente a la zona del espacio en que todas las coordenadas de un punto son no negativas.)

Por otra parte, tenemos una idea natural del grado de distinción de los colores. Por ejemplo, es claro que el rosa pálido está más próximo al blanco que el rosa fuerte, y el carmesí, más cerca del rojo que del azul, etc. Así, en el grado de distinción de los colores tenemos un concepto cualitativo de su distancia mutua. Este concepto cualitativo se puede transformar en medida cuantitativa. Sin embargo, definir la distancia entre colores como en la geometría euclidiana por la fórmula $r = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$ resulta artificial. Una distancia definida así no se corresponde con la percepción real; con esta definición ocurriría en muchos casos que dos colores que difieren de otro en distinto grado estarían a la misma distancia de él.

²⁶ Tales mezclas se pueden obtener combinando en distintas proporciones polvos muy finos coloreados, supuesto que la iluminación permanezca constante.

²⁷ Por ejemplo, si los colores C_0 y C_1 están determinados por sus intensidades, a saber, las coordenadas x_0, y_0, z_0 y x_1, y_1, z_1 , entonces un color C entre C_0 y C_1 tiene las coordenadas $x = (1 - t)x_0 + tx_1$, $y = (1 - t)y_0 + ty_1$, $z = (1 - t)z_0 + tz_1$, donde t es la proporción de C_1 , y $(1 - t)$, la de C_0 que entran en C .

La definición de distancia debe reflejar las relaciones reales entre las percepciones de color.

Guiados por este principio, introducimos una peculiar medida de distancia. Se efectúa como sigue. Al modificar un color continuamente, una persona no percibe el cambio sino cuando se supera el llamado umbral de distinción. En relación con esto, suponemos que todos los colores que están exactamente sobre el umbral de distinción de uno dado equidistan de él. Esto conduce automáticamente a la idea de que la distancia entre dos colores se debe medir por el menor número de umbrales que se pueden establecer entre ellos. La longitud de una línea de color se mide por el número de umbrales que la cubre. La distancia entre dos colores se mide por la longitud de la línea más corta que los une. Esto es similar a cómo se mide la distancia entre dos puntos en el plano.

Así, pues, la medida de longitudes y distancias en el espacio de color se lleva a cabo mediante pasos muy pequeños, infinitamente pequeños, como si dijésemos.

El resultado es que en el espacio de color se ha definido una cierta geometría no euclidiana que tiene un significado perfectamente real: describe en lenguaje geométrico las propiedades del conjunto de todos los colores posibles, es decir, las propiedades de las reacciones del ojo a los estímulos luminosos.

El concepto de espacio de color nació aproximadamente hace un siglo. Muchos son los físicos que han estudiado la geometría de este espacio; debemos mencionar, por ejemplo, a Helmholtz y Maxwell. Las investigaciones continúan y tienen un valor no sólo teórico, sino también práctico, pues dan una fundamentación matemática precisa a la solución de problemas sobre la diferencia de señales de color, sobre los tintes en la industria textil y otros.

2. *Espacios físicos en física y química*

Volvamos ahora a otro ejemplo ya mencionado en la sección precedente.

Supongamos que estamos estudiando un sistema fisicoquímico tal como una mezcla de gases, una aleación, etc. Supongamos además que el estado del sistema viene determinado por n valores (como el estado de una mezcla gaseosa lo está por la presión, la temperatura y las concentraciones de sus componentes). Se dice entonces que el sistema tiene n grados de libertad, queriendo decir con ello que su estado puede variar en n direcciones independientes al variar cada uno de los valores que lo determinan. Estos valores juegan el papel

Además, dados dos colores, pongamos rojo R y blanco B , si los mezclamos en proporciones variables ²⁶, obtenemos una gama de colores continua, desde el rojo R hasta el blanco B , que podemos llamar el segmento RB . La idea de que el color rosa está entre el rojo y el blanco tiene así un significado claro.

De este modo surge el concepto de las figuras y relaciones geométricas más sencillas en el «espacio de color». Un «punto» es un color, el «segmento» AB es el conjunto obtenido por la mezcla de los colores A y B ; la aserción «el punto D está en el segmento AB » significa que D es una mezcla de A y B . La mezcla de tres colores da un trozo de plano —un «triángulo de color». Todo esto se puede describir analíticamente usando las coordenadas de color x, y, z , y las fórmulas que dan las líneas de color y los planos de color son enteramente análogas a las de la geometría analítica ordinaria ²⁷.

En el espacio de color, las relaciones de la geometría de Euclides en lo que toca a la disposición de puntos y segmentos se satisfacen. El sistema de estas relaciones forma una geometría afín, y se puede decir que el conjunto de todas las posibles percepciones de color representan un espacio afín. (Esto no es completamente exacto, ya que, como hemos dicho, las coordenadas de color x, y, z no pueden ser negativas. Por tanto, el espacio de color corresponde solamente a la zona del espacio en que todas las coordenadas de un punto son no negativas.)

Por otra parte, tenemos una idea natural del grado de distinción de los colores. Por ejemplo, es claro que el rosa pálido está más próximo al blanco que el rosa fuerte, y el carmesí, más cerca del rojo que del azul, etc. Así, en el grado de distinción de los colores tenemos un concepto cualitativo de su distancia mutua. Este concepto cualitativo se puede transformar en medida cuantitativa. Sin embargo, definir la distancia entre colores como en la geometría euclidiana por la fórmula $r = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$ resulta artificial. Una distancia definida así no se corresponde con la percepción real; con esta definición ocurriría en muchos casos que dos colores que difieren de otro en distinto grado estarían a la misma distancia de él.

²⁶ Tales mezclas se pueden obtener combinando en distintas proporciones polvos muy finos coloreados, supuesto que la iluminación permanezca constante.

²⁷ Por ejemplo, si los colores C_0 y C_1 están determinados por sus intensidades, a saber, las coordenadas x_0, y_0, z_0 y x_1, y_1, z_1 , entonces un color C entre C_0 y C_1 tiene las coordenadas $x = (1 - t)x_0 + tx_1$, $y = (1 - t)y_0 + ty_1$, $z = (1 - t)z_0 + tz_1$, donde t es la proporción de C_1 , y $(1 - t)$, la de C_0 que entran en C .

La definición de distancia debe reflejar las relaciones reales entre las percepciones de color.

Guiados por este principio, introducimos una peculiar medida de distancia. Se efectúa como sigue. Al modificar un color continuamente, una persona no percibe el cambio sino cuando se supera el llamado umbral de distinción. En relación con esto, suponemos que todos los colores que están exactamente sobre el umbral de distinción de uno dado equidistan de él. Esto conduce automáticamente a la idea de que la distancia entre dos colores se debe medir por el menor número de umbrales que se pueden establecer entre ellos. La longitud de una línea de color se mide por el número de umbrales que la cubre. La distancia entre dos colores se mide por la longitud de la línea más corta que los une. Esto es similar a cómo se mide la distancia entre dos puntos en el plano.

Así, pues, la medida de longitudes y distancias en el espacio de color se lleva a cabo mediante pasos muy pequeños, infinitamente pequeños, como si dijésemos.

El resultado es que en el espacio de color se ha definido una cierta geometría no euclidiana que tiene un significado perfectamente real: describe en lenguaje geométrico las propiedades del conjunto de todos los colores posibles, es decir, las propiedades de las reacciones del ojo a los estímulos luminosos.

El concepto de espacio de color nació aproximadamente hace un siglo. Muchos son los físicos que han estudiado la geometría de este espacio; debemos mencionar, por ejemplo, a Helmholtz y Maxwell. Las investigaciones continúan y tienen un valor no sólo teórico, sino también práctico, pues dan una fundamentación matemática precisa a la solución de problemas sobre la diferencia de señales de color, sobre los tintes en la industria textil y otros.

2. *Espacios físicos en física y química*

Volvamos ahora a otro ejemplo ya mencionado en la sección precedente.

Supongamos que estamos estudiando un sistema fisicoquímico tal como una mezcla de gases, una aleación, etc. Supongamos además que el estado del sistema viene determinado por n valores (como el estado de una mezcla gaseosa lo está por la presión, la temperatura y las concentraciones de sus componentes). Se dice entonces que el sistema tiene n grados de libertad, queriendo decir con ello que su estado puede variar en n direcciones independientes al variar cada uno de los valores que lo determinan. Estos valores juegan el papel

de coordenadas. Por consiguiente, el conjunto de todos sus estados se puede considerar como un espacio n -dimensional; es lo que se llama el espacio fásico del sistema.

Los cambios continuos de estado, es decir, los procesos que experimenta el sistema, se representan mediante líneas en este espacio. Los distintos dominios de los estados, diferenciados por uno u otro rasgo, constituyen dominios del espacio fásico. Los estados que forman la frontera de dos de tales dominios constituyen una superficie en este espacio.

En la química física es particularmente importante estudiar la forma y contigüidad de aquellos dominios que corresponden a estados cualitativamente distintos. Las superficies que dividen dichos dominios corresponden a transiciones cualitativas tales como fusión, evaporación, precipitación de un sedimento, etcétera. En un sistema con dos grados de libertad, un estado se representa por un punto en un plano. Como ejemplo podemos tomar una substancia homogénea cuyo estado viene determinado por la presión p y la temperatura T ; éstas son las coordenadas que describen el estado. La cuestión se reduce entonces a estudiar las líneas de división entre dominios correspondientes a estados cualitativamente distintos. En el caso del agua, por ejemplo, los dominios son hielo, agua líquida y vapor (Fig. 30). Las líneas de división corresponden a la fusión (solidificación), evaporación (condensación) y sublimación del hielo (precipitación de cristales de hielo procedentes del vapor).

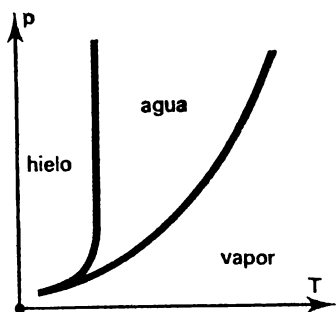


FIG. 30

Para el estudio de sistemas con mayor número de grados de libertad se requieren los métodos de la geometría multidimensional.

El concepto de espacio fásico se aplica no sólo a sistemas fisico-químicos, sino también a sistemas mecánicos, y en general se puede aplicar a cualquier sistema cuyos posibles estados formen un cierto conjunto continuo. En la teoría cinética de los gases, por ejemplo, se consideran los espacios fásicos de sistemas de partículas materiales: las moléculas de un gas. El estado de movimiento de una partícula está determinado en cada momento por su posición y velocidad, que

dan en conjunto 6 valores: tres coordenadas y tres componentes de la velocidad (con respecto a los tres ejes coordenados). El estado de N partículas está determinado por $6N$ valores, y puesto que su número es muy alto, $6N$ es un número enorme. Esto no preocupa a los físicos lo más mínimo, hablando así de un espacio fásico $6N$ -dimensional de un sistema de moléculas.

Un punto de este espacio describe el estado de toda la masa de moléculas con sus coordenadas y velocidad. El movimiento de un punto describe un cambio de estado. Esta presentación abstracta resulta muy conveniente en muchos desarrollos teóricos profundos. En una palabra, el concepto de espacio fásico tiene un lugar seguro en el arsenal de las ciencias naturales exactas y es aplicable a diversos problemas.

3. *La generalización de la geometría*

Los ejemplos dados permiten llegar a una conclusión acerca de cómo se debe generalizar el alcance de la geometría.

Supongamos que queremos estudiar cierta colección continua de objetos, fenómenos o estados de una u otra clase, como por ejemplo el conjunto de todos los colores posibles o de los estados de un grupo de moléculas. Puede suceder que las relaciones que prevalecen en una colección tal sean similares a las del espacio ordinario; por ejemplo, la «distancia» entre colores o la «posición mutua» entre dominios de un espacio fásico. En este caso, haciendo abstracción de las peculiaridades cualitativas de los objetos en cuestión y teniendo en cuenta solamente las susodichas relaciones, podemos considerar la colección dada como un espacio. Los «puntos» de este «espacio» son los propios objetos, fenómenos o estados. Una «figura» es un agregado arbitrario de sus puntos, como, por ejemplo, la «línea» de los colores del arco iris o el «dominio» de vapor en el «espacio» de los estados del agua. La «geometría» de tales espacios queda determinada por las relaciones «espaciales» que rigen entre los objetos, fenómenos o estados. Así, la geometría del «espacio de color» está determinada por las leyes de la mezcla de colores y por la distancia entre los mismos.

La importancia real de este punto de vista es que hace posible la utilización de los conceptos y métodos de la geometría abstracta en el estudio de diversos fenómenos. El campo de aplicación de los conceptos y métodos geométricos se amplía así sustancialmente. Como resultado de la generalización del concepto de espacio, el término «espacio» tiene en la ciencia dos acepciones: por una parte

es el espacio real ordinario (la forma universal de existencia de la materia), y por otra es el «espacio abstracto», una colección de objetos homogéneos (fenómenos, estados, etc.) entre los cuales existen ciertas relaciones de tipo espacial.

Es de notar que el espacio ordinario, tal y como lo visualizamos de una forma algo simplificada, se puede considerar también como una colección de estados homogéneos. A saber, como la colección de todas las posibles posiciones de un cuerpo infinitamente pequeño, es decir, un «punto material». Esta observación no pretende dar una definición del espacio, sino esclarecer la relación entre los dos tipos de espacio. El concepto de espacio abstracto será expuesto más ampliamente en la próxima sección, y la relación entre la geometría abstracta y el espacio real ordinario será tratada en la última sección de este capítulo.

4. *Espacios generalizados en matemáticas*

La aplicación más amplia del concepto de espacio abstracto tiene lugar en las propias matemáticas. En geometría consideramos el «espacio» de ciertas figuras, como, por ejemplo, el «espacio de las esferas» del cual ya hemos hablado, «el espacio de las líneas rectas», y así sucesivamente.

Este método se muestra extraordinariamente fructífero, en particular en la teoría de poliedros. Por ejemplo, en el §5 del capítulo 7 se mencionó un teorema sobre la existencia de un poliedro convexo con un desarrollo dado. La demostración del teorema se basa en la discusión de dos «espacios»: el «espacio de poliedros» y el «espacio de desarrollos». El conjunto de los poliedros convexos con un cierto número de vértices se considera como un espacio cuyos puntos representan poliedros; análogamente, el conjunto de los desarrollos se considera también como un espacio cuyos elementos representan desarrollos. El proceso de coordinar poliedros con desarrollos establece una correspondencia entre poliedros y desarrollos, es decir, entre los puntos del «espacio de poliedros» y los del «espacio de desarrollos». El problema consiste en probar que a cada desarrollo corresponde un poliedro; es decir, que a cada punto de un espacio corresponde un punto del otro. Y precisamente esto se puede demostrar aplicando métodos topológicos.

De forma similar se puede probar toda una serie de teoremas sobre los poliedros. El «método de espacios abstractos» resulta en algunos casos (como en el del teorema de existencia de un poliedro para un desarrollo dado) el más sencillo de todos los conocidos para

probar tales teoremas. Sin embargo, el método es lo suficientemente complicado como para que aquí no demos una explicación más detallada de él.

Aplicaciones extensivas del concepto de espacio generalizado se dan también en análisis, álgebra y teoría de números. Ello proviene de la usual representación de las funciones por medio de curvas. Los valores de la variable x se suelen dibujar como puntos de una recta. Análogamente, los valores de dos variables se representan por medio de puntos de un plano, los valores de n variables como puntos de un espacio n -dimensional; el conjunto de valores de las variables $x_1, x_2, \dots, x_n$ se representan por el punto de coordenadas $x_1, x_2, \dots, x_n$. Hablamos del «dominio de variación de las variables» o del «dominio de valores» de la función $f(x_1, x_2, \dots, x_n)$ de esas variables; hablamos de puntos, líneas o superficies, de discontinuidad de una función, etc. Este lenguaje geométrico es de uso constante y es algo más que un modo de expresión; la representación geométrica hace «intuitivos» muchos hechos del análisis por analogía con el espacio ordinario, y permite usar los métodos geométricos de demostración, generalizados al espacio n -dimensional.

Lo mismo ocurre en álgebra cuando se discuten ecuaciones con n incógnitas o funciones algebraicas de n variables. En la sección precedente hemos mencionado que una ecuación lineal con n incógnitas determina un hiperplano n -dimensional; que m de estas ecuaciones determinan m hiperplanos, y que cada solución del sistema representa un punto común a todos los hiperplanos. Los hiperplanos pueden no cortarse o cortarse en un solo punto, o en una línea recta, o en general en un hiperplano k -dimensional. En conjunto, los problemas sobre la solución de sistemas de ecuaciones lineales se interpretan como un problema de intersección de hiperplanos. Este enfoque geométrico tiene numerosas ventajas. El «álgebra lineal», que comprende el estudio de ecuaciones lineales y transformaciones lineales, se expresa muy a menudo en forma geométrica, como se hizo en el capítulo 15.

5. *El espacio de infinitas dimensiones; definición de un «espacio»*

En todos nuestros ejemplos hemos tratado con colecciones continuas de objetos de una u otra clase, considerándolas como espacios de una clase particular. Los objetos eran colores, estados de un sistema u otro, figuras o agregados de valores de variables. En todos los casos, los objetos venían dados por un número finito de datos, de modo que el espacio tenía un número finito de dimensiones, igual precisamente al número de datos.

Sin embargo, a comienzos del presente siglo, los matemáticos comenzaron a estudiar «espacios de infinitas dimensiones», a saber, colecciones de objetos que no se pueden determinar mediante un número finito de datos. Esto es lo que ocurre fundamentalmente en un «espacio funcional».

La idea de considerar una colección de funciones de uno u otro tipo como un espacio es una de las ideas básicas de una nueva rama del análisis, conocida como análisis funcional; y esta idea resulta extraordinariamente provechosa en la solución de muchos problemas. El lector encontrará en el capítulo 18, que está dedicado al análisis funcional, una exposición de esta disciplina.

Se puede hablar de espacios de funciones continuas de una o varias variables. También se pueden considerar como «espacios» distintas clases de funciones discontinuas para las que se define una distancia por uno u otro método, según lo aconseje la naturaleza del problema a resolver. En una palabra, el número de posibles «espacios funcionales» es ilimitado y muchos de ellos se emplean efectivamente en matemáticas.

Exactamente de la misma forma se puede hablar del «espacio de curvas», del «espacio de cuerpos convexos», del «espacio de los posibles movimientos de un sistema mecánico», etc. Por ejemplo, en el §5 del capítulo 7 mencionamos el siguiente teorema: sobre toda superficie convexa cerrada existen al menos tres geodésicas cerradas, y cada dos puntos se pueden unir por un número infinito de geodésicas. Para la demostración de este teorema se utiliza el espacio de curvas sobre la superficie y el espacio de curvas que unen dos puntos. En el conjunto de curvas que unen dos puntos dados introducimos un cierto tipo de distancia, obteniendo así un espacio. La demostración del teorema se basa en una aplicación de ciertos resultados profundos de la topología a este espacio.

Formulemos ahora una conclusión general.

Por «espacio» entendemos, de forma muy general, una colección arbitraria de objetos homogéneos (fenómenos, estados, funciones, figuras, valores de variables, etc.) entre los que existen relaciones similares a las relaciones espaciales usuales (continuidad, distancia, etcétera). Además, al considerar una colección de objetos como un espacio hacemos abstracción de todas las propiedades de los objetos a excepción de las determinadas por las relaciones en cuestión. Estas relaciones determinan lo que podemos llamar estructura o «geometría» del espacio. Los propios objetos juegan el papel de «puntos» del espacio; las «figuras» son conjuntos de «puntos».

El campo de la geometría de un espacio abstracto consiste en aquellas propiedades y figuras del espacio que están determinadas por las relaciones en cuestión. Por ejemplo, al estudiar el espacio de funciones continuas, las propiedades particulares de una función concreta son completamente ignoradas. La función juega el papel de un punto y por consiguiente «no tiene partes», no tiene en este sentido ninguna estructura, ni propiedades que estén desligadas de las demás funciones. En un espacio funcional, las propiedades de las funciones están determinadas solamente por sus relaciones con otras, por la distancia definida y por otras relaciones que se puedan derivar de la distancia.

A la variedad de posibles conjuntos de objetos y a las diversas relaciones entre ellos corresponde la ilimitada variedad de espacios que se estudian en matemáticas. Los espacios se pueden clasificar atendiendo a los distintos tipos de relaciones «espaciales» en que se basa su definición. Sin intentar dar una relación completa de los distintos tipos de espacio abstractos, hagamos mención, ante todo, de dos tipos muy importantes: los espacios topológicos y los métricos.

6. *Espacios topológicos*

Un *espacio topológico* (ver capítulo 17) es una colección de puntos (un conjunto arbitrario de elementos) en el que se ha definido una relación de proximidad de un punto a un conjunto y, en consecuencia, una relación de proximidad o adherencia entre dos conjuntos (figuras). Es una generalización de la relación de proximidad de figuras en el espacio ordinario.

Lobachevski señaló con notable perspicacia que la relación de proximidad era la más importante de todas las relaciones entre figuras. «La proximidad es una característica distintiva de los cuerpos que permite llamarlos *geométricos* cuando retenemos esta propiedad y prescindimos de las demás, sean accidentales o esenciales»²⁸. Por ejemplo, todo punto de la circunferencia es adherente al conjunto de los puntos interiores del círculo; dos partes de un cuerpo conexo son adherentes la una a la otra. Como ha demostrado el posterior desarrollo de la topología, es precisamente sobre la propiedad de proximidad o adherencia sobre la que se basan las demás propiedades topológicas.

El concepto de adherencia expresa la noción de que un punto está

²⁸ N. I. Lobachevski, *Obras completas*, Vol. II, Gostehizdat, 1949, pág. 168.

infinitamente próximo a un conjunto. Por tanto, toda colección de objetos en la que existe un concepto natural de continuidad, o de lo infinitamente próximo, es un espacio topológico.

El concepto de espacio topológico es extremadamente general, y el estudio de tales espacios —la topología abstracta— representa el estudio matemático más general de la continuidad.

Una definición rigurosa de espacio topológico general se puede dar de la siguiente forma.

Se dice que un conjunto arbitrario R de «puntos» es un espacio topológico general si para todo conjunto M contenido en él están definidos sus puntos adherentes, de suerte que se cumplan las condiciones siguientes, es decir, los axiomas del espacio.

1. Todo punto de M se cuenta entre sus puntos adherentes. (Es perfectamente natural suponer que cada punto de un conjunto es adherente a éste.)

2. Si un conjunto M_1 contiene a un conjunto M_2 , todos los puntos adherentes de M_2 lo son de M_1 . (De modo más breve, pero menos preciso: el conjunto mayor no debe tener menos puntos adherentes.)

A estos axiomas se suelen añadir otros, definiendo así los distintos tipos de espacios topológicos.

Con la ayuda del concepto de proximidad o adherencia es fácil definir una serie de conceptos topológicos muy importantes. Estos son, al mismo tiempo, los conceptos más fundamentales y generales de la geometría, y sus definiciones son intuitivamente muy claras. Demos algunos ejemplos.

1. *Conjuntos adherentes.* Decimos que los conjuntos M_1 y M_2 son adherentes si uno de ellos contiene al menos un punto adherente del otro. (En este sentido, por ejemplo, la circunferencia de un círculo es adherente al interior.)

2. *Continuidad* o, como se dice en matemáticas, *conexión* de una figura. Una figura, es decir, un conjunto de puntos M , se dice conexa si no se puede dividir en partes no adherentes entre sí. (Por ejemplo, un segmento es conexo, pero un segmento sin su punto medio no lo es.)

3. *Frontera.* La frontera de un conjunto M en un espacio R es el conjunto de los puntos adherentes tanto a M como a su complemento $R - M$, es decir, a la parte restante del espacio R . (Es, evidentemente, un concepto perfectamente natural de fronteras.)

4. *Punto interior.* Un punto de un conjunto M se llama interior si no pertenece a su frontera, es decir, si no es adherente a $R - M$.

5. *Aplicación o transformación continua.* Una transformación de un conjunto M se llama continua si no rompe las adherencias. (Difícilmente se podría dar una definición más natural de transformación continua.)

A esta lista cabría añadir otras definiciones importantes, como son la del concepto de convergencia de una sucesión de figuras a una figura dada, o la del concepto del número de dimensiones de un espacio.

Vemos, pues, que los conceptos geométricos más fundamentales se pueden definir en términos de adherencias. La importancia de la topología reside, en particular, en que es capaz de dar definiciones rigurosas generales de estos conceptos, proporcionando así una base para la estricta aplicación de los razonamientos relacionados con el concepto intuitivo de continuidad.

La topología es el estudio de aquellas propiedades de los espacios, figuras y transformaciones que vienen definidas por medio de la relación de adherencia.

La generalidad y naturaleza fundamental de esta relación hacen de la topología una teoría geométrica muy general que penetra las diversas ramas de las matemáticas donde quiera que lo único que interese sea la continuidad. Pero precisamente por su generalidad, la topología, en sus partes más abstractas, rebasa el marco de la propia geometría. Al mismo tiempo, en su base yace una generalización de las propiedades del espacio real, y sus resultados más fructíferos y potentes están relacionados con la aplicación de métodos que nacen de ideas geométricas intuitivas. Un ejemplo es el método de aproximación de figuras generales mediante poliedros, método que fue desarrollado por P. S. Aleksandrov y extendido por él, aunque en forma abstracta, a tipos muy generales de espacios topológicos.

Hoy en día cualquier especialista, no importa cuál sea su especialidad, cuando descubre que existe una forma natural de introducir los conceptos de proximidad o adherencia, tiene inmediatamente a su disposición el amplio aparato de la topología, lo cual le permite llegar a conclusiones nada triviales incluso en la aplicación a su especialidad.

7. Espacios métricos

Un *espacio métrico* es un conjunto de elementos arbitrarios entre los cuales se ha definido una distancia; es decir, a cada par de puntos X, Y se le asocia un número $r(X, Y)$ que satisfaga las siguientes condiciones o axiomas de espacio métrico:

1. $r(X, Y) = 0$ si, y sólo si, los puntos X e Y coinciden.
2. Para cada tres puntos X, Y, Z

$$r(X, Y) + r(Y, Z) \geq r(Z, X).$$

Esta condición se llama «desigualdad triangular», ya que es muy parecida a la conocida propiedad de la distancia ordinaria entre tres puntos, A_1, A_2, A_3 , del espacio euclidiano (Fig. 31):

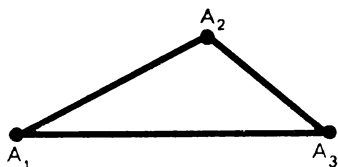


FIG. 31

$$r(A_1, A_2) + r(A_2, A_3) \geq r(A_3, A_1).$$

Como ejemplos de espacios métricos tenemos:

1. El espacio euclidiano de un número arbitrario n de dimensiones.
2. El espacio de Lobachevski.
3. Toda superficie con su métrica intrínseca (capítulo 7, §4).
4. El espacio C de las funciones continuas con la distancia definida por la fórmula $r(f_1, f_2) = \max |f_1(x) - f_2(x)|$.
5. El espacio de Hilbert que se describirá en el capítulo 18 y que es un «espacio euclidiano de infinitas dimensiones».

El espacio de Hilbert es el más importante de los utilizados en el análisis funcional; está muy relacionado con las series de Fourier y más generalmente con el desarrollo en serie de una función mediante funciones ortogonales (las coordenadas $x_1, x_2, x_3, \dots$, son los coeficientes de la serie). Este espacio también juega un importante papel en física matemática y es fundamental en la mecánica cuántica. También ocurre que el conjunto de todos los posibles estados (no sólo estacionarios) de un sistema atómico —por ejemplo, un átomo de hidrógeno— se puede considerar desde un punto de vista abstracto como un espacio de Hilbert.

El número de ejemplos de espacios métricos que actualmente se estudian en matemáticas se podría aumentar considerablemente; en la próxima sección nos familiarizaremos con una importante clase de espacios métricos, los llamados espacios riemannianos. Pero los ejemplos dados ahora bastarán para mostrar cuán amplio es el concepto de espacio métrico.

En un espacio métrico se pueden definir, además de todos los conceptos topológicos, otros conceptos métricos; por ejemplo, el concepto de longitud de una curva en un espacio métrico: la longitud se define exactamente de la misma forma que en el espacio usual, conservando así las propiedades básicas de la longitud. Efectivamente, por longitud de una curva entendemos el límite de la suma de las distancias entre puntos $X_1, X_2, \dots, X_n$ que forman una sucesión sobre la curva, cuando la distancia entre dichos puntos se hace cada vez más pequeña.

8. *Ventajas del método geométrico*

Aparte de los espacios topológicos y métricos hay muchos otros espacios que se estudian en matemáticas. De hecho, en el §6 hemos tenido ocasión de examinar toda una clase de tales espacios: son los espacios en los que se da un cierto grupo de transformaciones (por ejemplo, los espacios afines y proyectivos). En tales espacios se puede definir la «igualdad» de figuras. Dos figuras son «iguales» si se puede llevar la una sobre la otra mediante una transformación del grupo dado.

No profundizaremos más en las definiciones de posibles tipos de espacios; son muy diversos y el lector interesado puede consultar la literatura especial sobre las distintas ramas que la geometría tiene hoy día.

¿Pero qué sentido tiene extender de tal forma el campo de los conceptos geométricos? ¿Para qué, por ejemplo, nos hace falta introducir el concepto de espacio de funciones continuas? ¿No son suficientes los métodos usuales para resolver los problemas de análisis sin recurrir a tales espacios abstractos?

La respuesta general a estas cuestiones consiste, brevemente, en lo siguiente: mediante la introducción de uno u otro espacio en el estudio abrimos el camino a la aplicación de los conceptos y métodos geométricos, los cuales son extremadamente numerosos.

Un hecho característico de los conceptos y métodos geométricos es que se basan, *grosso modo*, en ideas intuitivas y conservan sus ventajas incluso en su forma abstracta. Lo que el analista obtiene

después de largos cálculos el geómetra puede alcanzarlo en ocasiones inmediatamente. Un ejemplo elemental de esto se puede ver en las gráficas, que dan una imagen completa de la dependencia entre las variables. El método geométrico se puede caracterizar como un método sintético que lo abarca todo, en contraste con el método analítico. Ciertamente es que en las teorías geométricas abstractas el carácter intuitivo inmediato se desvanece, pero los razonamientos intuitivos analógicos permanecen, al igual que el carácter sintético del método geométrico.

El lector está ya familiarizado con la aplicación de las imágenes geométricas en análisis, con la representación gráfica de los números complejos y de las funciones de una variable compleja, con razonamientos geométricos en una de las demostraciones del teorema fundamental del álgebra y con otras aplicaciones de los conceptos y métodos geométricos. Por todas partes se observa lo que hemos descrito aquí en términos generales. Recordemos los ejemplos al comienzo del §7, y también aquí en el §8, en la subsección 5), de teoremas que se pueden probar aplicando la geometría multidimensional. Demos un ejemplo más de un problema de análisis que se puede resolver aplicando el concepto de espacio funcional.

En topología se prueba que, si tomamos en el plano ordinario un dominio que tenga la forma de un círculo deformado y le aplicamos una deformación continua, de manera que al final quede contenido dentro de su contorno original, entonces existe al menos un punto que al final de la deformación queda donde estaba al principio. Se trata de un resultado puramente topológico.

Consideremos un problema completamente ajeno a la geometría: se trata de encontrar una función $y(x)$ que satisfaga la ecuación diferencial ²⁹,

$$y' = f(x, y) \quad (10)$$

y tal que para $x = 0$ tome el valor $y = 0$.

Evidentemente, en lugar de esta ecuación, podemos considerar el problema de buscar una solución de la ecuación

$$y = \int_0^x f[t, y(t)] dt. \quad (11)$$

Naturalmente, surge el problema: ¿existe, en general, una función $y(x)$ que satisfaga esta condición?

²⁹ Suponemos que $f(x, y)$ es una función continua de las variables x e y .

Observemos el problema de otra forma. Representemos cada función continua como un punto de un espacio abstracto. El resultado del cálculo de la integral

$$\int_0^x f[t, y(t)] dt = z(x)$$

es una función continua de x , es decir, un «punto» de nuestro espacio abstracto. Tomando distintos «puntos» y , es decir, funciones $y(x)$, obtenemos, en general, distintos puntos z . De esta forma se establece una correspondencia entre los puntos del espacio. El problema de encontrar una solución del problema (11) se reduce a la cuestión: ¿existe un punto en nuestro espacio que, después de transformado, coincida con su «lugar» previo?

Un problema natural de la teoría de ecuaciones diferenciales se convierte así en un problema relacionado con cierta propiedad de un espacio abstracto de funciones. La analogía con el teorema antes mencionado nos dice que, evidentemente, estamos ante una propiedad topológica de dicho espacio.

De esta forma, por medio de los procedimientos topológicos necesarios se obtienen las demostraciones probablemente más cortas de muchos teoremas de existencia de soluciones de ecuaciones diferenciales; en particular, se demuestra fácilmente que la ecuación (10) tiene, de hecho, una solución para toda función $f(x, y)$ continua.

§9. Geometría riemanniana

1. Historia de la geometría riemanniana

La idea, expuesta anteriormente, de que toda colección continua de fenómenos homogéneos puede considerarse como un espacio fue expresada por primera vez por Riemann en su conferencia «Sobre las hipótesis en que se basa la geometría», pronunciada en la Universidad de Gotinga en 1854. Se trataba de la disertación o conferencia que todo profesor tenía que pronunciar en la facultad antes de tomar posesión de su cátedra. En esta conferencia Riemann expuso en líneas generales, sin cálculos ni demostraciones matemáticas, la idea original de una vasta teoría geométrica que hoy se denomina geometría riemanniana. Se dice que nadie le comprendió en el auditorio excepto el anciano Gauss. Riemann redactó el aparato formal de su teoría en otro escrito, con una aplicación a la conducción del

calor, de modo que la geometría riemanniana abstracta nació muy ligada a la física matemática. En el desarrollo de la geometría, las ideas de Riemann aparecieron inmediatamente después del paso decisivo de Lobachevski. Sin embargo, sus ideas no fueron debidamente apreciadas. Sus conferencias y trabajos fueron publicados después de su muerte en el año 1868. Es de señalar que la primera interpretación de la geometría de Lobachevski apareció también en 1868, obra de Beltrami, y la segunda, dada por Klein, en 1872. En 1872 Klein expuso su punto de vista general sobre las distintas geometrías: euclidiana, de Lobachevski, proyectiva, afín, etc., considerando las como el estudio de aquellas propiedades de las figuras que permanecen inalteradas en las transformaciones de ciertos grupos (de transformaciones). En el mismo año se consolidó la geometría multidimensional. Así, pues, los años setenta del siglo XIX constituyen el período crítico de la historia de la geometría en el que las ideas geométricas, acumuladas en el curso de los cincuenta años precedentes, fueron por fin comprendidas por un amplio grupo de matemáticos, consiguiendo así un lugar seguro en la ciencia.

El trabajo de Riemann fue continuado, y a finales del siglo XIX la geometría riemanniana había alcanzado un desarrollo considerable y encontrado aplicaciones en mecánica y en física. Cuando, en 1915, Einstein aplicó la geometría riemanniana (en su teoría general de la relatividad) a la teoría de la gravitación universal, la atención se centró en dicha geometría, originando su rápido desarrollo y dando lugar a diversas aplicaciones a distintos campos de la ciencia.

2. *Las ideas básicas de la geometría de Riemann*

Las ideas de Riemann, que tan brillante éxito han conocido, resultan en realidad muy simples si se dejan a un lado los detalles matemáticos, concentrándose en los fundamentos esenciales. Tal simplicidad intrínseca es una característica de las grandes ideas. ¿No fue simple la idea de Lobachevski de considerar las consecuencias de la negación del Quinto Postulado como una posible geometría? ¿No fue simple la idea de la evolución de los organismos, o la de la estructura atómica de la materia? Todas ellas son simples y al mismo tiempo complicadas, ya que las nuevas ideas deben, en primer lugar, abrirse camino en un amplio campo y no confinarse en un sistema rígido, y en segundo lugar, su fundamentación, desarrollo y aplicación es una tarea múltiple que requiere un inmenso acopio de labor e ingenio y que resulta imposible sin los métodos especializados de la ciencia. El aparato de la geometría riemanniana consiste en sus

fórmulas; son complicadas y por tanto accesibles sólo al especialista. Pero nosotros no nos pararemos en estas fórmulas, sino que iremos a la esencia de las ideas de Riemann. Como ya hemos dicho, Riemann comenzó considerando una colección continua de fenómenos homogéneos como un espacio. En este espacio las coordenadas de los puntos son cantidades que determinan el fenómeno en cuestión, como por ejemplo, las intensidades x, y, z determinan el color $C = xR + yV + zA$. Si hay n valores —pongamos $x_1, x_2, \dots, x_n$ — entonces hablamos de un espacio n -dimensional. En este espacio podemos considerar líneas e introducir una medida de su longitud en pequeños pasos (infinitamente pequeños), de modo similar a como se mide la longitud de una curva en el espacio ordinario.

Para medir longitudes en pasos infinitamente pequeños, basta dar una regla que determine la distancia de un punto dado a otro infinitamente próximo a él. Esta regla de determinación (medida) de distancias recibe el nombre de *métrica*. El caso más simple se presenta cuando la regla es la misma que en el espacio euclidiano. Tales espacios son euclidianos en lo infinitamente pequeño. En otros términos, la relaciones geométricas de la geometría euclidiana se satisfacen en él, pero solamente en dominios infinitamente pequeños; es más exacto, aunque tampoco del todo, decir que estas relaciones se satisfacen en dominios suficientemente pequeños con una aproximación tanto mayor cuanto menor es el dominio. Un espacio en el que la distancia se mide por una regla de este tipo se llama *riemanniano*; y la geometría de tales espacios se llama también *riemanniana*. Un espacio riemanniano es por tanto un espacio que es euclidiano «en lo infinitamente pequeño».

El ejemplo más sencillo de espacio riemanniano es una superficie lisa cualquiera en su geometría intrínseca. La geometría intrínseca de una superficie es una geometría riemanniana de dos dimensiones. Pues, en el entorno de cada uno de sus puntos, una superficie lisa difiere muy poco de su plano tangente y esta diferencia es tanto más pequeña cuanto menor es el dominio considerado. Por tanto, en un dominio pequeño de la superficie, su geometría difiere también poco de la geometría del plano; cuanto más pequeño es el dominio, menores son las diferencias. Sin embargo, en dominios mayores la geometría de una superficie curva es diferente de la de Euclides (como ya explicamos en el §4 del capítulo 7 y como es fácil de ver en los ejemplos de la esfera o pseudoesfera). La geometría riemanniana no es más que una generalización natural de la geometría intrínseca de una superficie de dos dimensiones a un número arbitrario n . Igual que

en una superficie, un espacio riemanniano tridimensional, considerado solamente desde el punto de vista de su geometría intrínseca, no deferirá apenas del euclidiano en dominios pequeños, pero sí en dominios extensos. Por ejemplo, la longitud de una circunferencia puede no ser proporcional al radio; sólo será proporcional (con una buena aproximación) para circunferencias pequeñas. La suma de los ángulos de un triángulo puede no ser dos rectos; el papel de segmentos rectilíneos en la construcción de un triángulo lo juegan aquí las líneas de mínima distancia, es decir, las líneas que entre todas las que unen los puntos dados tienen la mínima longitud.

Podemos decir que el espacio real es euclidiano solamente en dominios que son pequeños en comparación con la escala astronómica. Cuanto más pequeño es un dominio más aproximadamente se cumple la geometría euclidiana, pero podemos imaginar (y así ocurre) que a escala muy grande la geometría del espacio difiere algo de la euclidiana. Esta idea, como ya sabemos, fue señalada por Lobachevski. Riemann la generalizó aplicándola a una geometría arbitraria y no sólo a la de Lobachevski; esta última aparece como un caso especial de la geometría riemanniana.

De lo que llevamos dicho se deduce que la geometría riemanniana nació como síntesis de tres ideas que contribuyeron al feliz desarrollo de la geometría. La primera fue la idea de la posibilidad de una geometría distinta de la euclidiana; la segunda, el concepto de geometría intrínseca de una superficie, y la tercera, el concepto de espacio con un número arbitrario de dimensiones.

3. *Medida de la distancia*

Para comprender cómo se define matemáticamente un espacio riemanniano, vamos a recordar en primer lugar la regla para medir distancias en un espacio euclídeo.

Supuestas unas coordenadas rectangulares x, y en el plano, por el teorema de Pitágoras la distancia entre dos puntos cuyas diferencias de coordenadas son Δx e Δy se expresa por la fórmula

$$s = \sqrt{\Delta x^2 + \Delta y^2}.$$

Análogamente, en un espacio tridimensional tenemos:

$$s = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}.$$

En un espacio euclídeo n -dimensional, la distancia se define por la fórmula general

$$s = \sqrt{\Delta x_1^2 + \cdots + \Delta x_n^2}.$$

De aquí es fácil deducir la regla para medir distancias en un espacio riemanniano. La regla debe coincidir con la euclidiana, pero solamente para dominios infinitamente pequeños en el entorno de cada punto. Esto conduce al siguiente enunciado de la regla.

Un espacio riemanniano se caracteriza por el hecho de que en el entorno de cada uno de sus puntos A se pueden introducir unas coordenadas $x_1, x_2, \dots, x_n$, de modo que la distancia de A a un punto X infinitamente próximo venga expresada por la fórmula

$$ds = \sqrt{dx_1^2 + \cdots + dx_n^2}, \quad (12)$$

donde $dx_1, \dots, dx_n$ son las diferencias (infinitamente pequeñas) entre las coordenadas de A y X . Esto se puede expresar más precisamente de la forma siguiente: la distancia de A a un punto arbitrariamente próximo X se expresa por la misma fórmula que en la geometría euclidiana, pero solamente con una cierta aproximación, que es tanto cuanto más próximos son los puntos X y A , es decir,

$$s(AX) = \sqrt{\Delta x_1^2 + \cdots + \Delta x_n^2} + \epsilon,$$

donde ϵ es una cantidad pequeña en comparación con el primer término y es tanto más pequeña cuanto menores sean las diferencias de coordenadas $\Delta x_1, \dots, \Delta x_n$.³⁰

Esta es la definición matemática precisa de métrica riemanniana y de espacio riemanniano. La diferencia entre la métrica riemanniana (es decir, la regla para medir distancias) y la euclidiana consiste en que esta regla es válida solamente en el entorno de cada punto. Además las coordenadas en las que se expresa de un modo tan simple son diferentes para cada punto³¹. La diferencia entre la métrica

³⁰ El significado preciso de la fórmula (12) se suele expresar como sigue. Supongamos que una curva parte de A y que las coordenadas de un punto X arbitrario de ella vienen dados por las funciones $x_1(t), x_2(t), \dots, x_n(t)$ de una variable t . Entonces la diferencial ds de la longitud de un arco de curva en A viene expresada por (12).

³¹ Si se pudieran introducir en todo el espacio unas coordenadas tales que esta regla para medir distancias se cumpliera para cada par de entornos, entonces el espacio sería euclidiano.

riemanniana general y la euclidiana será estudiada con detalle más adelante.

El hecho de que un espacio riemanniano coincide con el euclídeo en lo infinitamente pequeño permite definir en él las magnitudes geométricas fundamentales, de modo similar a como hicimos para la geometría intrínseca de una superficie mediante la aproximación de una parte infinitamente pequeña de ella por un plano (capítulo 7, §4). Por ejemplo, un volumen infinitamente pequeño se expresa igual que en el espacio euclídeo. El volumen de un dominio finito se obtiene sumando volúmenes infinitamente pequeños, es decir, integrando la diferencial de volumen. La longitud de una curva se obtiene sumando distancias infinitamente pequeñas entre puntos infinitamente próximos de ella, es decir, por integración de la diferencial de longitud ds a lo largo de la curva. Y esto es una expresión analítica rigurosa del hecho de que la longitud se determina llevando sobre la curva una vara de medir infinitamente pequeña. El ángulo entre dos curvas en un punto común se determina igual que en el espacio. Además, en un espacio riemanniano n -dimensional se pueden definir superficies de distintas dimensiones, desde 2 hasta $n - 1$. Por otra parte, es fácil probar que cada superficie, a su vez, representa un espacio riemanniano del correspondiente número de dimensiones, de la misma forma que una superficie en el espacio euclídeo resulta ser un espacio riemanniano bidimensional.

También se demuestra que un espacio de Riemann se puede representar siempre como una superficie en un espacio euclidiano de un número de dimensiones suficientemente grande, a saber: para cada espacio riemanniano n -dimensional se puede encontrar en un espacio euclídeo $n(n + 1)/2$ -dimensional una superficie que, desde el punto de vista de su superficie intrínseca, no difiera de este espacio riemanniano (al menos en una parte de él).

4. *La forma cuadrática fundamental*

Para obtener la expresión analítica efectiva de las distintas cantidades geométricas, tenemos que definir, en primer lugar, una expresión general para la regla de medir longitudes en un espacio de Riemann, de modo que dicha expresión sea independiente de las coordenadas específicas de cada punto. Ciertamente, la fórmula (12) se verifica en cada punto A para una elección apropiada de coordenadas en dicho punto, de forma que al pasar a otro punto es preciso cambiar las coordenadas, lo cual es, desde luego, un inconveniente. Pero esto se puede evitar fácilmente y para ello probaremos lo siguiente.

Supongamos que en un cierto dominio de un espacio de Riemann se introducen unas coordenadas arbitrarias $y_1, y_2, \dots, y_n$. Entonces «la distancia infinitamente pequeña» o, como se suele decir, el «elemento de longitud» desde un punto A de coordenadas $y_1, y_2, \dots, y_n$ a otro punto X de coordenadas $y_1 + dy_1, y_2 + dy_2, \dots, y_n + dy_n$ se expresa por la fórmula

$$ds = \sqrt{\sum_{i,k=1}^n g_{ik} dy_i dy_k}, \quad \text{o} \quad ds^2 = \sum g_{ik} dy_i dy_k, \quad (13)$$

donde los coeficientes g_{ik} son funciones de las coordenadas $y_1, y_2, \dots, y_n$ de A .

La expresión de la derecha es una forma cuadrática³² en las diferenciales de las coordenadas $y_1, y_2, \dots, y_n$. Desarrollándola se puede escribir así:

$$\sum g_{ik} dy_i dy_k = g_{11} dy_1^2 + g_{12} dy_1 dy_2 + g_{21} dy_2 dy_1 + g_{22} dy_2^2 + \dots$$

Puesto que $dy_1 dy_2 = dy_2 dy_1$, es conveniente tomar el segundo y tercer término iguales: $g_{12} = g_{21}$, y en general $g_{ik} = g_{ki}$; esto es posible, ya que solamente su suma $(g_{ik} + g_{ki}) dy_i dy_k$ es importante.

La forma cuadrática es definida positiva, ya que evidentemente $ds^2 > 0$ excepto cuando todas las diferenciales son nulas.

La inversa también es cierta. A saber, si en un espacio n -dimensional, en el que se han introducido las coordenadas $y_1, y_2, \dots, y_n$, se define el elemento de longitud por la fórmula (13) con la condición de que la fórmula cuadrática sea definida positiva (es decir, siempre mayor que 0 excepto cuando todas las $dy = 0$), entonces el espacio es riemanniano. En otros términos, en el entorno de cada punto A se pueden introducir nuevas coordenadas $x_1, x_2, \dots, x_n$ de modo que en ellas el elemento de longitud se expresa en la forma sencilla (12):

$$ds^2 = dx_1^2 + dx_2^2 + \dots + dx_n^2.$$

Así, pues, la métrica riemanniana (es decir, una definición de longitud que es euclidiana en lo infinitamente pequeño) se puede dar mediante cualquier forma cuadrática definida positiva de la forma (13)

³² Una forma cuadrática de varias variables es una expresión algebraica que es un polinomio de grado dos en esas variables.

con coeficientes g_{ik} que sean funciones de las coordenadas y_i . Este es el método general de dar una métrica riemanniana.

En un espacio de Riemann, una curva viene dada por la variación de las n -coordenadas de un punto respecto a un solo parámetro t que toma valores en un cierto intervalo

$$y_1 = y_1(t), y_2 = y_2(t), \dots, y_n = y_n(t) \quad (a \leq t \leq b). \quad (14)$$

La longitud de la curva se expresa mediante la integral

$$s = \int ds = \int \sqrt{\sum g_{ik} dy_i dy_k}.$$

En el caso de que la curva esté dada por las ecuaciones (14) tenemos

$$dy_1 = y'_1 dt, \dots, dy_n = y'_n dt,$$

por tanto

$$s = \int_a^b \sqrt{\sum g_{ik} y'_i y'_k} dt. \quad (15)$$

Puesto que las g_{ik} son funciones conocidas de las coordenadas $y_1, \dots, y_n$, y estas últimas dependen de t de acuerdo con las fórmulas (14), la función de t bajo el signo integral está completamente determinada para la curva dada. En consecuencia, la integral tiene un valor definido y la curva tiene también una longitud definida.

La longitud de la curva más corta que une dos puntos dados A y B se toma como la distancia entre estos puntos. Esta curva, llamada geodésica, juega un papel análogo al del segmento rectilíneo AB . Se puede demostrar que en un dominio pequeño cada dos puntos están unidos por una única geodésica. El problema de hallar una línea geodésica es el de hallar el mínimo de la integral (15). Es un problema de cálculo de variaciones que se discutió ya en el capítulo 8. La aplicación del método del cálculo de variaciones permite deducir una ecuación diferencial, que determina las líneas geodésicas, y establecer sus propiedades para todo espacio de Riemann.

Probemos una de las proporciones dadas anteriormente, a saber, que la métrica riemanniana en coordenadas arbitrarias viene dada mediante la fórmula general (13).

Supongamos que en un dominio de un espacio riemanniano se introducen ciertas coordenadas $y_1, y_2, \dots, y_n$. Tomamos un punto A

arbitrario en este dominio y suponemos que $x_1, x_2, \dots, x_n$ son las coordenadas especiales en las cuales el elemento de longitud en A se expresa mediante la fórmula (12) o, lo que es lo mismo,

$$ds^2 = dx_1^2 + dx_2^2 + \cdots + dx_n^2. \quad (16)$$

Las coordenadas x_j se expresan en función de las $y_i (i, j = 1, \dots, n)$ por ciertas fórmulas,

$$\begin{aligned} x_1 &= f_1(y_1, y_2, \dots, y_n), \\ x_2 &= f_2(y_1, y_2, \dots, y_n), \\ &\vdots \\ x_n &= f_n(y_1, y_2, \dots, y_n). \end{aligned}$$

Entonces

$$dx_1 = \frac{\partial f_1}{\partial y_1} dy_1 + \frac{\partial f_2}{\partial y_2} dy_2 + \cdots + \frac{\partial f_n}{\partial y_n} dy_n$$

y análogamente, para $dx_2, \dots, dx_n$. Sustituimos estas expresiones en (16). Después de elevar al cuadrado y agrupar los términos en $dy_1^2, dy_1 dy_2, dy_2^2$, etc., obtenemos una expresión de la forma

$$ds^2 = g_{11} dy_1^2 + 2g_{12} dy_1 dy_2 + g_{22} dy_2^2 + \cdots + g_{nn} dy_n^2,$$

donde los coeficientes $g_{11}, g_{12}, \dots, g_{nn}$ son expresiones en las derivadas parciales $\partial f_i / \partial y_j$ (la forma de estas expresiones no nos interesa ahora). Pero esto no es más que la fórmula (13) escrita en forma desarrollada, y por consiguiente la proposición está probada.

Demostremos que, inversamente, la fórmula (13) define una métrica riemanniana, es decir, que en cada punto se puede transformar, con una elección adecuada de las coordenadas x_i , en la fórmula (16). Supongamos que

$$ds^2 = \sum g_{ik} dy_i dy_k,$$

donde las g_{ik} son funciones de las coordenadas $y_1, \dots, y_n$ y la forma cuadrática de la derecha es definida positiva. Entonces los coeficientes g_{ik} son números dados y las variables de las que depende la forma son $dy_1, \dots, dy_n$. Sabemos por álgebra que toda forma definida positiva se puede reducir a una suma de cuadrados mediante un

cio riemanniano general y otro euclidiano mediante la definición de una medida, por así decirlo, de la no-euclidianidad del espacio riemanniano. Esta medida es lo que se llama tensor de curvatura del espacio.

Debemos subrayar ante todo que el concepto de curvatura de un espacio no está relacionado en absoluto con la idea de que el espacio esté sumergido en un espacio superior envolvente en el cual tenga, de algún modo, una cierta curvatura. La curvatura se define dentro del espacio dado y expresa su diferencia respecto al euclidiano en el sentido de sus propiedades geométricas intrínsecas. Esto debe quedar muy claro con el fin de no relacionar el concepto de espacio curvado con asuntos extraños. Cuando se dice que el espacio real tiene curvatura, esto significa solamente que sus propiedades geométricas difieren de las del espacio euclidiano. Pero esto no significa en absoluto que nuestro espacio esté inmerso en otro espacio en el que sea, de alguna manera, curvo. Tal idea no tiene relación alguna con las aplicaciones de la geometría riemanniana al espacio real y pertenece al campo de la fantasía especulativa.

El concepto de curvatura de un espacio de Riemann generaliza a n dimensiones el de curvatura gaussiana de una superficie. Como ya explicamos en el §4 del capítulo 7, la curvatura gaussiana es una medida de la diferencia entre la geometría intrínseca de una superficie y la geometría en un plano y se puede tratar desde un punto de vista puramente geométrico interno. Y no es otra cosa que la curvatura del espacio riemanniano bidimensional que representa la superficie dada.

Recordemos, por ejemplo, dos fórmulas de la geometría intrínseca en las que interviene la curvatura de Gauss. Imaginemos un pequeño triángulo sobre la superficie, próximo a un punto O y cuyos lados son líneas geodésicas; sean sus ángulos α, β, γ y su área σ . La cantidad $\alpha + \beta + \gamma - \pi$ expresa la diferencia entre la suma de sus ángulos y la suma de los ángulos de un triángulo en el plano.

Cuando el triángulo se contrae hacia el punto O , la razón de $\alpha + \beta + \gamma - \pi$ a su área σ tiende a la curvatura gaussiana en O . De otra forma, para un pequeño triángulo

$$\frac{\alpha + \beta + \gamma - \pi}{\sigma} = K + \epsilon,$$

donde $\epsilon \rightarrow 0$ cuando el triángulo se contrae hacia O . Esto muestra exactamente que la curvatura gaussiana es una medida de la dife-

rencia entre la suma de los ángulos de un triángulo en el plano y en la superficie.

Consideremos ahora un pequeño círculo sobre la superficie con centro en O (es decir, el lugar geométrico de los que equidistan de O , en el sentido de la distancia *sobre* la superficie). Si r es el radio del círculo y l su longitud, entonces

$$l = 2\pi r - \frac{\pi}{3} Kr^3 + \epsilon,$$

donde K es otra vez la curvatura gaussiana en O y ϵ denota una cantidad muy pequeña en comparación con r^3 .

Aquí la curvatura gaussiana aparece como una medida de la diferencia entre la longitud de un pequeño círculo y $2\pi r$, que es su longitud en la geometría euclidiana.

La curvatura de un espacio riemanniano juega un papel similar. Se puede definir, por ejemplo, de la siguiente manera. En un espacio reimanniano dado construimos una superficie lisa F formada por geodésicas que pasan por un punto O . Tomamos como curvatura del espacio en O , en la dirección de F , la curvatura gaussiana de esta superficie. En general, esta curvatura dependerá no sólo del punto O , sino también de las distintas superficies geodésicas G que pasan por O . La curvatura de un espacio en un punto dado no está, por tanto, caracterizada por un solo número. El propio Riemann introdujo una regla general para relacionar las curvaturas de las distintas superficies en un mismo punto. Debido a estas relaciones, la curvatura en un punto está completamente caracterizada por un cierto sistema de números llamado *tensor de curvatura*.

Sin embargo, no nos detendremos aquí en un desarrollo de la situación, ya que ello requiere un extenso aparato matemático. Lo únicamente importante es darse cuenta de que el tensor de curvatura es una medida de la no euclidianidad de un espacio riemanniano y está definido intrínsecamente como una medida de la diferencia entre su métrica y la del espacio euclidiano. Determina, por ejemplo, la diferencia entre la suma de los ángulos de un triángulo y π , o entre la longitud de una circunferencia y $2\pi r$. En puntos distintos tiene, en general, valores distintos y en el mismo punto viene dado no por un número, sino por un cierto sistema de números.

Un espacio riemanniano no es necesariamente homogéneo en sus propiedades y en este caso es imposible mover libremente las figuras sin alterar la distancia entre sus puntos. Interesa entonces saber en

qué espacios riemannianos es posible un movimiento libre de las figuras con el mismo número de grados de libertad que en el espacio euclidiano. Estos son los espacios riemannianos más homogéneos.

El espacio euclidiano es un espacio homogéneo sin curvatura (un espacio de curvatura cero). Otro tipo de espacio homogéneo es el de Lobachevski, de modo que la geometría de Lobachevski, exactamente igual que la de Euclides, es un caso especial de la geometría riemanniana.

En general, un espacio de Riemann en el que es posible el movimiento libre de figuras es un espacio de curvatura constante: la curvatura es la misma para todos los puntos y para todas las superficies geodésicas. (En lugar de un «tensor de curvatura» que cambia de un punto a otro, ahora viene dada por un único número común a todos los puntos.) Un espacio de curvatura cero es euclidiano; un espacio de curvatura negativa es un espacio de Lobachevski; un espacio de curvatura positiva tiene la misma geometría que una esfera n -dimensional en un espacio euclidiano $(n + 1)$ -dimensional.

6. Aplicaciones de la geometría riemanniana

La geometría riemanniana no necesitó esperar mucho para encontrar aplicaciones. El mismo Riemann, como ya hemos dicho, aplicó el aparato formal de esta geometría a la solución de un problema sobre la conducción del calor, pero fue una mera aplicación de sus fórmulas y no de la idea de espacio abstracto con una métrica euclidiana en dominios infinitamente pequeños. Tal aplicación fue hecha en el espacio de color, donde la distancia entre colores próximos se expresó utilizando una métrica riemanniana; el espacio de color ha sido tratado como un espacio riemanniano tridimensional.

Otra importante aplicación de la geometría riemanniana se llevó a cabo en mecánica. Para comprender su esencia consideremos, para comenzar, el movimiento de un punto material sobre una superficie. Imaginemos un punto material, por ejemplo, una partícula que se mueve libremente sobre una superficie sin abandonarla. El punto se mueve, por así decirlo, en la superficie misma. Introduzcamos unas coordenadas arbitrarias x_1, x_2 sobre la superficie; entonces el movimiento del punto está completamente determinado por la dependencia de esas coordenadas respecto al tiempo, y su velocidad por la velocidad de cambio de las coordenadas, es decir, por las derivadas con respecto al tiempo $\dot{x}_1, \dot{x}_2$. Vemos, pues, que los puntos se mueven, por así decirlo, en un espacio bidimensional; pero este espacio no es

euclidiano y tiene su propia geometría, la geometría intrínseca de la superficie. Las leyes del movimiento se pueden transformar de modo que contengan solamente las coordenadas x_1, x_2 de un punto sobre la superficie, y sus primeras y segundas derivadas.

Si sobre el punto una fuerza actúa, entonces su componente normal a la superficie se anula con la reacción de ésta, quedando solamente las componentes tangenciales a la superficie. De esta forma las fuerzas que actúan sobre un punto se pueden considerar como actuando en la misma superficie. La geometría intrínseca de la superficie es un caso especial de una geometría riemanniana. Por tanto, el movimiento de un punto sobre la superficie es un movimiento en un espacio riemanniano bidimensional. Las leyes de este movimiento tienen el mismo carácter que las leyes usuales, con la única diferencia de que se toma en cuenta la geometría de la superficie. Esto se ve perfectamente claro en el siguiente hecho, mencionado ya anteriormente en el §4 del capítulo 7: un punto que se mueve por inercia sobre una superficie sin rozamiento describe una geodésica con velocidad constante. Puesto que las líneas geodésicas juegan sobre la superficie el papel de líneas rectas, este hecho es análogo a la ley de inercia: es la misma ley de inercia, pero referida al movimiento sobre una superficie, o abstractamente, en un espacio riemanniano bidimensional.

Desde luego, hasta aquí no se ve por ningún lado la ventaja de esta presentación abstracta, ya que hemos tratado solamente del movimiento sobre una superficie ordinaria.

Las ventajas del punto de vista abstracto se perciben en el acto cuando se aplica a un sistema mecánico cuyos estados están determinados por más de dos cantidades, pues entonces resulta imposible representar el movimiento como el movimiento de un punto sobre una superficie. Con esta situación hemos topado ya en el §7 cuando hablábamos de cómo los métodos gráficos fracasaban en la representación de un espacio multidimensional.

Supongamos entonces que tenemos un sistema mecánico cuya configuración, es decir, la distribución de sus partes, está determinada por n cantidades $x_1, x_2, \dots, x_n$. Si lo que estamos estudiando es un sistema de varios puntos materiales, su distribución vendrá determinada por todas sus coordenadas, tres para cada punto. Podemos tomar como otro ejemplo un giroscopio (un disco girando alrededor de un eje que a su vez se mueve alrededor de un punto fijo). La rotación del giroscopio alrededor del eje se da por el ángulo de deflexión, y la inclinación del eje, por los dos ángulos que forma con dos direc-

ciones dadas. En total son tres las cantidades que determinan la posición del giroscopio (Fig. 32).

Cada configuración se puede considerar como un «punto» del espacio de todas las configuraciones posibles. Este es el espacio llamado espacio de configuración del sistema³⁵. Su dimensión es igual al número de cantidades $x_1, x_2, \dots, x_n$ que determinan la configuración. Estas cantidades sirven de coordenadas de un «punto» en el espacio de configuración. Para un sistema de tres puntos materiales, por ejemplo, tenemos tres coordenadas para cada punto y nueve en total. Para el caso del giroscopio tenemos tres coordenadas (los tres ángulos), de modo que su espacio de configuración es tridimensional.

El movimiento del sistema se representa como el movimiento de un punto en un espacio de configuración. La velocidad del movimiento está determinado por la velocidad de cambio de las coordenadas $x_1, x_2, \dots, x_n$.

Tales espacios serán discutidos en el capítulo 17 en relación con su estructura topológica. Aquí queremos señalar solamente que en un espacio de configuración se puede introducir una regla especial para medir distancias que esté estrechamente relacionada con las propiedades mecánicas del sistema. Efectivamente, si la energía cinética del sistema se expresa por la fórmula

$$T = \frac{1}{2} \sum_{i,k=1}^n a_{ik} \dot{x}_i \dot{x}_k,$$

donde las $\dot{x}_i$ son las velocidades de cambio de las correspondientes coordenadas, entonces el cuadrado de una distancia infinitamente pequeña viene dado por la fórmula

$$ds^2 = \sum_{i,k=1}^n a_{ik} dx_i dx_k.$$

³⁵ No debe confundirse con el «espacio fásico» mencionado en §8 en la subsección 2. En el espacio fásico un punto determina no sólo la posición, sino también la velocidad de movimiento de los puntos del sistema en cada momento.

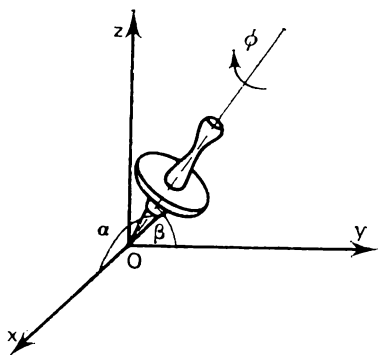


FIG. 32

Así, el espacio de configuración se convierte en un espacio riemanniano. Por otro lado, aparte de que el movimiento del sistema viene representado por el movimiento de un punto en el espacio de configuración, también las ecuaciones que describen el movimiento del sistema coinciden con las ecuaciones del movimiento de este punto; en una palabra, la mecánica del sistema está representada por la mecánica de un punto en el espacio de configuración. En particular, el movimiento inercial del sistema, es decir, en ausencia de fuerzas (como el movimiento libre de rotación del giroscopio) da lugar a un movimiento uniforme del punto sobre una línea geodésica en ese espacio.

Esta representación es muy conveniente en muchos casos y se utiliza, con ciertas modificaciones y generalizaciones, en mecánica teórica.

Así, pues, la geometría riemanniana tiene aplicaciones como método de representación abstracto-geométrica de fenómenos físicos. Esta descripción no es en absoluto arbitraria, y tampoco es un juego ocioso de la mente matemática; refleja el mecanismo real de los fenómenos en cuestión, aunque lo hace de forma abstracta. Ahora bien, esa es la naturaleza de todas las descripciones matemáticas de los fenómenos físicos y esa es también la naturaleza de toda aplicación de la geometría abstracta, con la única diferencia de que las abstracciones son aquí más potentes y delicadas. Pero la esencia es la misma.

La aplicación más brillante de la geometría riemanniana se produjo en la teoría de la relatividad. De esto hablaremos en la próxima sección, donde trataremos el importante y difícil problema de las relaciones entre la geometría abstracta y el espacio real.

7. Generalizaciones

En los últimos treinta años la geometría de diversos espacios no euclidianos ha sido objeto de importantes desarrollos y generalizaciones en distintas direcciones. Han surgido nuevas teorías en las que la geometría riemanniana se incluye como caso especial. La primera fue la llamada geometría de Finsler, cuya idea se remonta directamente hasta Riemann³⁶; vino luego una teoría general de espacios, obra del eminente geómetra francés E. Cartan, en la cual se combina la geometría riemanniana con el Erlanger Programm de Klein, y otras teorías. No nos es posible dar una exposición detallada de estas

³⁶ Finsler fue el matemático alemán que inició en 1916 un estudio detallado de la geometría que lleva su nombre.

nuevas direcciones de la geometría, y mencionaremos tan sólo que han sido obtenidas esencialmente por medio de especiales adaptaciones de su aparato analítico. Un grupo de geómetras soviéticos ha tomado parte en el desarrollo de estas nuevas direcciones; podríamos nombrar quizá la nueva geometría «polimétrica» creada por P. K. Rasovski y los estudios de V. V. Vagner que se extienden desde los problemas más generales de la teoría de espacios curvados a las aplicaciones de la geometría no euclidiana a la mecánica.

§10. *La geometría abstracta y el espacio real*

1. *Dificultades en la visualización del espacio real como no euclidiano*

En el curso de la precedente exposición del desarrollo de las ideas geométricas, comenzando por Lobachevski, hemos ahondado en los espacios abstractos, alejándonos del objeto original de la geometría: el espacio real en el cual tienen lugar todos los fenómenos. Volveremos ahora a este espacio, imponiéndonos la tarea de describir cómo el desarrollo de la geometría abstracta ha contribuido al conocimiento de sus propiedades.

Sabemos que la geometría nació del experimento, de una investigación empírica de las formas espaciales y de las relaciones entre los cuerpos: de la medida de parcelas de tierra, de volúmenes de recipientes, etc., Es, pues, en su origen, una teoría física como lo es, pongamos por caso, la mecánica. Los axiomas de la geometría euclidiana eran conclusiones claramente formuladas sobre la base de prolongados experimentos; expresan leyes de la naturaleza y podrían llamarse leyes de la geometría, de la misma forma que las leyes de la mecánica se llaman a menudo axiomas de la mecánica³⁷. Pero es equivocado afirmar que estas leyes son absolutamente exactas y que jamás requerirán modificación o generalización alguna en relación con los nuevos datos experimentales; las propiedades del espacio pueden diferir más o menos de lo que afirma la geometría de Euclides.

³⁷ Durante los últimos cincuenta años ha surgido una concepción abstracta de los axiomas que los disocia de su contenido original; pero esto no cambia en nada el hecho de que los axiomas de la geometría euclidiana expresan hechos de la naturaleza. Cuando hablamos de axiomas y no de leyes de la geometría o de la mecánica queremos poner de relieve la construcción lógico-deductiva de estas ciencias, pero por ello no pierden su carácter experimental. Un enunciado de un teorema se llama axioma cuando se toma como base para la construcción deductiva de la teoría, deduciendo otros enunciados de la teoría a partir de los básicos (axiomas) por razonamientos lógicos.

Nosotros hemos expuesto ya estos argumentos y creemos que ahora deben aparecer muy evidentes. Pero no ocurría lo mismo en aquellos días en que las ideas de Lobachevski tropezaban aún con la incomprensión general. Antes de Lobachevski y Gauss, a nadie se le ocurrió pensar que la geometría de Euclides pudiera no ser enteramente exacta ni que las propiedades reales del espacio pudieran diferir de aquélla. Lobachevski desarrolló su geometría como una teoría de las posibles propiedades del espacio real. Más tarde Riemann y otros científicos plantearon también el problema de las posibles propiedades del espacio, de las posibles leyes para medir longitudes que se podrían descubrir por procedimientos más aproximados. En general, cada geometría abstracta, en alguna de sus partes, se puede considerar como una teoría de posibles propiedades del espacio. Todo esto perteneció, no obstante, al campo de las hipótesis hasta que en 1915 Einstein, en su teoría general de la relatividad, corroboró las ideas de Lobachevski y Riemann. Esta teoría deja ver claramente que la geometría del espacio real difiere de hecho un poco de la de Euclides, y esto se descubrió precisamente sobre la escala astronómica, cosa que Lobachevski había ya anticipado.

En lo que acabamos de decir acerca del espacio se vislumbran, por lo menos, tres dificultades. El problema de la relación de la geometría abstracta con la geometría física, es decir con las propiedades del espacio real, se reduce en último término al esclarecimiento de esas dificultades.

La primera de ellas consiste en concebir cómo y en qué sentido las propiedades del espacio real pueden diferir de las afirmaciones de la geometría euclidianas. Estamos tan acostumbrados a ella que no podemos imaginar nada distinto, y por tanto algunas explicaciones son evidentemente necesarias.

La segunda dificultad reside en la propia expresión «propiedades del espacio real». El espacio en sí mismo se concibe como algo vacío y homogéneo. Podría parecer que en el concepto mismo de espacio va ya incluida la idea de su homogeneidad. ¿Pero cómo puede ser que el espacio vacío, es decir, la «vaciedad», tenga alguna propiedad? Hablamos de las «propiedades del espacio» sin pensar en estos problemas, pero merece la pena pararse en ellos, como se deduce de las dificultades mencionadas.

La tercera dificultad está en la verdad de una u otra geometría. La cuestión puede parecer muy simple: es verdadera aquella geometría que corresponde a la realidad. Así es, desde luego. Pero por otra parte hemos visto, por ejemplo, que la geometría en el interior de un

círculo se puede considerar como una geometría de Lobachevski, ya que todo hecho geométrico dentro de un círculo se puede presentar como un teorema de esta última geometría. En consecuencia, resulta que los mismos hechos geométricos se pueden presentar como teoremas de ambas geometrías, la de Euclides y la de Lobachevski. Por tanto, ambas geometrías corresponden a la realidad. Entonces, ¿cuál de ellas es verdadera y en qué sentido, y por qué suponemos, a pesar de todo, que la geometría de Euclides se satisface en el círculo y que la de Lobachevski está solamente ilustrada o interpretada en él?

Evidentemente, estos problemas encierran una dificultad considerable que ha llegado a desconcertar incluso a algunos matemáticos eminentes.

Nuestras explicaciones deben comenzar por la segunda de las dificultades mencionadas, ya que una comprensión de lo que son las «propiedades del espacio» nos conducirá a la solución de las otras dos preguntas.

2. *Espacio y materia*

El objeto de la geometría, a saber, las «propiedades del espacio», está integrado por las propiedades de los cuerpos reales, sus relaciones materiales y formas. En el espacio real, un «lugar», un «punto», una «dirección», son cosas que vienen determinadas por los cuerpos materiales. «Aquí» y «allí», «hacia aquí» y «hacia allá», sólo tienen significado en relación con uno u otro objeto material. «Aquí» puede significar «sobre la tierra» o «en esta habitación» o alguna cosa por el estilo; en una palabra, «aquí» significa siempre un lugar determinado por un determinado criterio material. De modo similar, una línea recta no existe por sí misma, sino solamente como una cuerda tensa, o como la arista de una regla, o como un rayo de luz. Una línea recta, «una línea como ésta», es una abstracción que refleja la propiedad común de esas líneas materiales del mismo modo que, por ejemplo, la «casa en sí misma» es una abstracción que refleja las propiedades comunes de todas las casas; la «casa en sí misma» no existe fuera o independientemente de las diversas casas reales.

Este carácter objetivo de las propiedades del espacio está expresado en el conocido principio del materialismo dialéctico: *El espacio es la forma de existencia de la materia*. La forma de un objeto está determinada por las conexiones y relaciones entre sus partes. La estructura del espacio es la regularidad común de una serie de relaciones de cuerpos materiales y fenómenos. Están las relaciones espaciales, el orden espacial de los objetos, su posición mutua, distancias, etcé-

tera. Pero como la forma es inseparable de su contenido, excepto en abstracciones y en ciertos contextos, así también el espacio es inseparable de la materia. La idea de un espacio «en sí mismo», de un espacio sin materia, es una abstracción de la que no debemos abusar. Las relaciones y formas espaciales: «aquí», «entre», «dentro», «recto», «esfera», etc., son siempre relaciones y formas de los cuerpos materiales. La geometría, sin embargo, los considera abstractamente. Esta abstracción es necesaria, porque de otra forma no sería posible percibir ninguna generalidad en las diversas relaciones concretas entre los objetos. Pero a esta abstracción no debe elevarse a la categoría de absoluta mediante una sustitución de la propia realidad objetiva por conceptos abstractos.

En un espacio absolutamente vacío, carente de toda traza de materia, nada podría distinguir un lugar, un punto, una dirección, y en consecuencia no habría ni lugares, ni puntos, ni direcciones, de modo que el espacio absolutamente vacío se reduce a nada. Incluso en la idea abstracta de espacio vacío suponemos tácitamente que en él se pueden distinguir distintos lugares y direcciones. Dicho con otras palabras, en la idea abstracta de espacio conservamos las propiedades de distinción de lugares, direcciones y distancias, que existen en el espacio real precisamente porque este espacio está inseparablemente ligado a los cuerpos materiales.

Así, pues, el espacio es la forma de existencia de la materia; las «propiedades del espacio» son, por tanto, propiedades de la materia, propiedades de ciertas relaciones de cuerpos materiales, sus posiciones mutuas, dimensiones, etc.

Además, si queremos que los teoremas de la geometría tengan un significado físico, debemos saber qué entendemos por «línea recta», «distancia» y otros conceptos geométricos. En la sección 4 hemos visto que una misma teoría geométrica admite distintas interpretaciones.

En consecuencia, para contrastar la geometría con la experiencia es necesario definir tan exactamente como sea posible el significado físico de los conceptos geométricos, ya que la geometría describe las propiedades del espacio real, pero sólo a condición de que a sus conceptos se les asigne el significado físico correspondiente. Sin este significado físico los teoremas de la geometría son una abstracción matemática de carácter formal. Aquí reside la solución de la segunda dificultad antes mencionada. Esta dificultad surge porque en lugar de pensar en el espacio real, inseparablemente ligado a la materia, uno intenta concebir un espacio «puro», un espacio «en sí mismo», que, sin embargo, no es más que una abstracción.

3. *Intuición y comprensión*

Es fácil comprender ahora cómo se resuelven las otras dos dificultades. En primer lugar, ¿cómo podemos imaginar que el espacio real es, en cuanto a sus propiedades, distinto del euclidiano? Supongamos que queremos verificar un resultado concreto de la geometría euclidiana, por ejemplo, que la suma de los ángulos de un triángulo es 180° , o que la longitud de una circunferencia es $2\pi R$. Para verificar lo primero debemos determinar qué triángulos físicos se van a considerar y cómo se definen sus ángulos. Supongamos que un lado de un triángulo es un rayo de luz en el espacio vacío. En este caso no hay nada inconcebible en el hecho de que experimentos muy exactos lleguen a demostrar que la suma de los ángulos de un triángulo no sea 180° . De la misma forma, cabe imaginar que las medidas de una circunferencia y de su radio con una misma escala no satisfagan la relación $l = 2\pi R$ exactamente. De hecho así ocurre en la superficie de la tierra, donde la longitud de la circunferencia no es proporcional al radio, sino que crece lentamente y alcanza un máximo cuando el radio es igual a un semimeridiano. Se puede objetar, no obstante, que sobre la superficie de la tierra el papel de las rectas lo juegan los arcos de círculos máximos, de modo que el radio se entiende en otro sentido, y nuestro resultado no contradice por tanto a la geometría euclidiana. Sin embargo, de acuerdo con la teoría de la relatividad, cerca de un cuerpo de gran masa la razón de la longitud de la circunferencia a su radio «genuino» es, a pesar de todo, diferente de 2π ; de hecho se verifica la siguiente fórmula aproximada para la razón de la longitud del ecuador de un cuerpo homogéneo en forma de esfera a su radio:

$$\frac{\text{longitud de la circunferencia}}{\text{radio}} = 2\pi \left(1 - \frac{kM}{3Rc^2} \right),$$

donde M es la masa del cuerpo en toneladas métricas, R el radio del cuerpo en kilómetros, $c = 300.000$ km/seg la velocidad de la luz, y k la constante de gravitación, que en nuestras unidades vale $66,6 \cdot 10^{-18}$.

Vemos entonces que la razón entre la longitud de una circunferencia y su radio es un poco menor que 2π . Los cálculos muestran que sobre la superficie del sol la razón difiere de 2π en aproximadamente 0,000004, y sobre la superficie del compañero de Sirio, cuya densidad media es 50.000 veces la del agua, la diferencia alcanza un valor de 0,00014.

Se puede objetar, desde luego, que todo esto es algo imposible de imaginar y que en una visión intuitiva el espacio será siempre euclidiano. Esta objeción no debe preocuparnos porque, en primer lugar, la tarea de la ciencia no consiste en dar imágenes intuitivas de los fenómenos, sino en llegar a una comprensión de ellos. Toda imagen intuitiva está limitada y condicionada por las formas usuales transmitidas por los órganos sensoriales. Por tanto, no estamos en condiciones de tener una idea intuitiva de los rayos ultravioleta, ni de la propagación de radioondas, ni del movimiento del electrón en un átomo, ni de muchos otros fenómenos, si no es sustituyéndolos por modelos. Pero esto no significa que todos estos fenómenos sean incomprensibles para nosotros. Por el contrario, el éxito de la radio-tecnología, por ejemplo, demuestra que hemos llegado a dominar las radioondas y que por tanto las comprendemos muy bien. En segundo lugar, la solución del problema de lo que podemos y no podemos imaginar depende del hábito y entrenamiento de la imaginación. ¿Podemos imaginarnos a los antípodas que, desde nuestro punto de vista, andan cabeza abajo? Hoy sí que podemos imaginarlos, pero hubo un tiempo en que la «inimaginabilidad» de los antípodas sirvió de argumento en contra de la redondez de la tierra.

4. *Geometría y verdad*

Volvamos ahora a la última dificultad, es decir, al problema de saber qué geometría debe considerarse como verdadera. Cuando planteamos el problema, indicamos que los hechos geométricos dentro del círculo se podían considerar como teoremas de la geometría de Euclides y de la de Lobachevski. Por tanto, ambas geometrías corresponden a la realidad; es decir, ambas son verdaderas. Y después de todo, no hay nada de sorprendente en ello. Un mismo fenómeno se puede describir siempre por varios métodos; una misma cantidad se puede medir con distintas unidades; una misma curva puede ser descrita mediante distintas ecuaciones, según sean las coordenadas elegidas. Análogamente, una cierta colección de relaciones geométricas (en el ejemplo en cuestión, las relaciones dentro del círculo) se puede describir por distintos métodos. Pero nosotros planteamos el problema, no de una colección aislada de hechos geométricos, sino de las relaciones espaciales en toda su generalidad. El espacio es la forma universal de existencia de la materia y, consecuentemente, cuando se plantea el problema de las propiedades del espacio no se puede segregar artificialmente un dominio concreto de hechos.

Planteando el problema de esta forma, no se pueden separar ya las cantidades o los hechos geométricos de otros fenómenos a los que van necesariamente unidos. Por ejemplo, la longitud de un segmento se determina transportando sobre él una regla rígida de forma que la medida de longitudes está necesariamente relacionada con el movimiento de cuerpos rígidos, es decir, con la mecánica. La geometría es inseparable de la mecánica. Sin embargo, la medida de longitud dentro de un círculo en la interpretación de la geometría de Lobachevski se lleva a cabo de manera muy diferente, como ya explicamos en el §4; una cuerda posee aquí una longitud infinita. Está claro que la medida así definida no se corresponde con la idea original de medida, nacida sobre la base del transporte mecánico de los cuerpos a comparar. En general, las figuras que son iguales en el sentido de la geometría de Euclides son, por el mismo origen de la geometría euclidiana, figuras que se pueden superponer mediante un movimiento mecánico. En la interpretación de la geometría de Lobachevski, la igualdad se define de modo distinto, y el papel de movimientos lo asumen otras transformaciones. Por tanto, cuando los hechos geométricos dentro del círculo se consideran junto con sus necesarias ligazones con la mecánica, debemos admitir que es en efecto la geometría de Euclides la que se verifica dentro del círculo (con gran aproximación).

La geometría euclidiana es la única en la cual el papel de movimiento lo desempeña el movimiento mecánico ordinario de los cuerpos rígidos. Esta es precisamente la razón de que fuese la geometría euclidiana, y no otra geometría, la primera en ser descubierta. Pero el desarrollo de la física nos ha llevado ahora a la conclusión de que las leyes de la mecánica newtoniana, y con ellas las de la geometría euclídea, son sólo aproximaciones de otras más exactas y generales. En este cambio de las leyes de la geometría, en el paso de la geometría euclídea a la riemanniana, utilizada luego en la teoría de la relatividad, han influido no sólo la mecánica, sino también otras ramas de la física: de igual, si no mayor, importancia fueron la teoría de los fenómenos electromagnéticos y la óptica. La geometría como ciencia de las propiedades del espacio está relacionada con la física, depende de ella y solamente puede ser separada de ésta en la abstracción y en ciertos contextos.

La dependencia de la geometría respecto a la física, o en otras palabras, la dependencia de las propiedades del espacio respecto a la materia, fue claramente señalada por Lobachevski, quien previó la posibilidad de un cambio en las leyes de la geometría en relación con

los nuevos descubrimientos físicos. En contraste con este punto de vista materialista, el famoso matemático Poincaré afirmó hace algún tiempo que la elección de una u otra geometría estaba dictada solamente por motivos de simplicidad o «economía de pensamiento», en la frase del conocido idealista Mach. Sobre esta base, Poincaré aseguró que se abandonaría antes la ley de la propagación rectilínea de la luz que la geometría euclidiana, ya que ésta es «la más simple». sin embargo, Poincaré murió tres años antes de que la teoría de la relatividad fuese finalmente desarrollada; y en ella sucede precisamente lo contrario: se abandona la geometría euclidiana, conservando en cambio la ley de propagación de la luz, aunque en forma generalizada: la luz se propaga según líneas geodésicas.

Hemos llegado así a las siguientes conclusiones. Una misma colección aislada de hechos se puede describir, en general, de distintas formas, y todas esas descripciones son verdaderas, siempre que reflejen la realidad. Sin embargo, es equivocado considerar los hechos geométricos —en su entera generalidad— como divorciados de otros fenómenos. Solamente así se pueden establecer las propiedades del espacio, ya que éste es la forma universal de existencia de la materia. Pero al tomar la geometría en conjunción con la física, es preciso adaptar la una a la otra. Es entonces cuando se pone de manifiesto la diferencia esencial entre las distintas «geometrías», las cuales, cuando se disocian de la física, sólo se pueden distinguir por su mayor o menor simplicidad. La geometría euclidiana apareció no porque fuera la más simple, sino porque se correspondía con la mecánica. De hecho, en relación con el desarrollo de la física en la teoría de la relatividad, pasamos a otra geometría más complicada, que es la riemanniana.

Resumiendo: en relación a las propiedades del espacio real, es verdadera aquella geometría que las refleja en su entera generalidad con suficiente aproximación y, en consecuencia, da cuenta no sólo de los hechos puramente geométricos, sino también de los de la mecánica y física en general.

5. *El espacio-tiempo de la relatividad*

Lo poco que llevamos dicho sobre la teoría de la relatividad no atañe a su contenido principal. En la comprensión de los problemas del espacio fue mucho más allá de lo que Lobachevski y Riemann hubieran podido pensar.

La proposición más esencial y básica de la teoría de la relatividad es la siguiente: el espacio es completamente inseparable del tiempo, y

ambos forman una única forma de existencia de la materia, la variedad tetradimensional del espacio-tiempo. Un suceso del universo se caracteriza por la posición y por el tiempo y en consecuencia por sus cuatro coordenadas: tres espaciales y la cuarta temporal, el tiempo del suceso. Los sucesos forman en este sentido una colección tetradimensional. La teoría de la relatividad se interesa ante todo por esta colección, dejando de lado las propiedades de los fenómenos individuales. No es fundamentalmente una teoría de movimientos rápidos, ni de la cosmología ni una nueva teoría del espacio o del tiempo, sino precisamente una teoría del espacio-tiempo como forma única de existencia de la materia.

Desde luego, en la mecánica newtoniana también se pueden combinar el espacio y el tiempo en una única variedad tetradimensional. Como vimos ya, la idea de un espacio multidimensional apareció en un trabajo de Lagrange cuando, al considerar el movimiento de un punto material, añadió a las coordenadas espaciales x, y, z la coordenada del tiempo t . El movimiento de un punto se representa entonces como una línea en el espacio tetradimensional con las coordenadas x, y, z, t ; al moverse el punto, las cuatro coordenadas varían: la posición (x, y, z) y el tiempo t . Sin embargo, esta unificación del espacio y el tiempo tiene un carácter puramente formal. En ello no se pone de manifiesto ninguna relación interna y necesaria entre el espacio y el tiempo. Por supuesto, en la ley del movimiento de un cuerpo existe una dependencia entre la posición espacial y el tiempo. Pero esto concierne solamente a cada movimiento concreto; ni en mecánica ni en física general se había establecido antes de la teoría de la relatividad una relación universal entre el espacio y el tiempo. Las relaciones espaciales, el orden espacial de los objetos y fenómenos fueron siempre cuidadosamente distinguidos de sus relaciones y orden en el tiempo. La sucesión temporal de sucesos, la duración de los intervalos de tiempo se consideraban como absolutos, como independientes de lo que sucediera en ellos. En resumen, privaba el concepto de tiempo absoluto.

El mayor descubrimiento de Einstein, que fue no sólo la piedra angular de la teoría de la relatividad, sino que revolucionó toda la física y la comprensión filosófica del problema del espacio y tiempo, fue el descubrimiento de que el tiempo absoluto no existe. Poco después de que Einstein desarrollara su teoría en 1905³⁸, Minkowski mostró

³⁸ La teoría que Einstein desarrolló en 1905 recibe el nombre de teoría especial de la relatividad, en oposición a la teoría «general» de 1915.

que su esencia consiste no tanto en el rechazo del tiempo absoluto como en el establecimiento de una ligazón entre el espacio y el tiempo en virtud del cual existe una única forma absoluta de existencia de la materia: el espacio-tiempo. La separación entre el espacio (las coordenadas espaciales) y el tiempo (la coordenada del tiempo t) es hasta cierto punto relativa y depende del sistema material (el «sistema de referencia») en relación al cual se determina el orden espacial y temporal de los fenómenos. Sucesos que son simultáneos con referencia a un sistema no tienen por qué serlo con referencia a otro.

La definición del orden de los fenómenos no puede ser, por supuesto, completamente dependiente del sistema de referencia. El orden de sucesos que están relacionados por una interacción directa es el mismo, naturalmente, en todos los sistemas de referencia, de modo que una acción siempre precede a su resultado. Pero, para sucesos que no están relacionados por una interacción, su orden en el tiempo es relativo. Puesto que el orden espacial (en su forma pura) se refiere a sucesos simultáneos y la simultaneidad es relativa, resulta que la separación de las relaciones puramente espaciales respecto de la colección general de relaciones espacio-tiempo es también relativa y dependiente del sistema de referencia. El espacio, en sentido abstracto, es, por decirlo así, una «sección» de la variedad tetradimensional del espacio-tiempo en la que están contenidos todos los sucesos simultáneos (en relación a un sistema dado).

Nuestro objetivo no es desarrollar los fundamentos de la teoría de la relatividad, por lo cual nos limitaremos a caracterizar en breves palabras sus rasgos básicos y la forma más natural en que deben ser considerados en relación con las ideas de la geometría abstracta. Esta interpretación, dicho sea de paso, es muy diferente de la que dio el propio Einstein.

El mundo, el universo, se puede contemplar como un conjunto de diversos sucesos. Entendemos aquí por suceso no un fenómeno arbitrario que se extiende en el espacio y en el tiempo, sino, por así decirlo, un fenómeno instantáneo y puntual, un «flash» momentáneo de una fuente puntual de luz. Utilizando el lenguaje geométrico: los sucesos son puntos en la variedad tetradimensional del universo.

El espacio-tiempo es la forma de existencia de la materia, la forma de esta variedad universal. La estructura del espacio-tiempo, su «geometría», no es más que una cierta estructura general del universo, es decir, de acuerdo con nuestro análisis, la «geometría» del conjunto de sucesos. Esta estructura está determinada por ciertas conexiones y relaciones materiales y universales de los sucesos.

En primer lugar, como ya hemos explicado para las relaciones espaciales, se trata de relaciones y conexiones *materiales*. Lo mismo ocurre con las relaciones de los fenómenos en el tiempo. Las relaciones espaciales y temporales, «como tales», en forma pura, son sólo abstracciones.

En segundo lugar, las relaciones (entre sucesos) que determinan la estructura del espacio-tiempo deben tener carácter universal, de acuerdo con el carácter universal del espacio-tiempo.

Tal relación material universal es la conexión causa-efecto. Todo suceso actúa, en una forma u otra, directamente sobre otros sucesos y experimenta a su vez la acción de otros. Esta relación de acción de unos sucesos sobre otros determina la estructura del espacio-tiempo.

Así, la teoría de la relatividad nos permite dar la siguiente definición: el espacio-tiempo es el conjunto de todos los sucesos, sin tener en cuenta las propiedades y relaciones concretas y considerando solamente la relación general de acción de unos sucesos sobre otros. Esta relación debe entenderse aquí en un sentido también general, independientemente de sus distintas formas concretas.

En la teoría especial de la relatividad, el espacio-tiempo se considera como totalmente homogéneo. Esto significa que la variedad de los sucesos admite transformaciones que conservan las relaciones de acción entre los sucesos, y que el grupo de estas transformaciones es, en cierto sentido, lo más amplio posible. Supongamos, por ejemplo, dos pares de sucesos A, B y A', B' y que ni A actúa sobre B ni B sobre A , e igualmente para A', B' . Existe entonces una transformación entre los sucesos, según la cual A se corresponde con A' y B con B' , tal que para cada par de sucesos se conserva la relación de acción (o inacción); es decir, si X actúa sobre Y , entonces el correspondiente suceso X' también actúa sobre Y' ; y si X no actúa sobre Y , tampoco X' actúa sobre Y' .

De acuerdo con estas consideraciones, resulta que, desde el punto de vista de la teoría espacial de la relatividad, el propio espacio-tiempo es un espacio tetradimensional cuya geometría está determinada por un cierto grupo de transformaciones. Estas transformaciones no son otras que las famosas transformaciones de Lorentz. Las leyes de la geometría y la física no varían bajo estas transformaciones. Este punto de vista corresponde a la visión de la geometría expuesta por Klein en su «Erlanger Programm», del cual ya hablamos en la sección 6.

6. *Gravitación y curvatura*

La teoría general de la relatividad va más allá y abandona la idea de la homogeneidad del espacio-tiempo. Supone que el espacio-tiempo es homogéneo sólo con una cierta aproximación en dominios suficientemente pequeños, pero que no lo es en su totalidad. De acuerdo con la teoría de Einstein, la no homogeneidad del espacio está determinada por la distribución y movimiento de materia. Por su parte, la estructura del espacio-tiempo determina las leyes de movimiento de los cuerpos, y esto se manifiesta en el fenómeno de la gravitación universal. La teoría general de la relatividad es, en sentido propio, una teoría de la gravitación que explica el nexo gravitatorio de la estructura del espacio-tiempo con el movimiento de la materia.

La idea de considerar el espacio-tiempo como homogéneo sólo con una cierta aproximación en dominios pequeños es similar a la idea de que un espacio riemanniano es euclidiano sólo en «lo infinitamente pequeño». El espacio-tiempo matemático de la teoría de la relatividad se maneja como una clase de espacio riemanniano, aunque en un sentido sustancialmente distinto.

De hecho, en un espacio riemanniano tetradimensional podemos introducir unas coordenadas tales en el entorno de cada punto que el cuadrado de la longitud de un elemento de línea venga expresado por la fórmula $ds^2 = dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2$.

En el espacio-tiempo se pueden introducir coordenadas x, y, z, t en el entorno de cada suceso de forma que la longitud del elemento de línea venga dado por la fórmula $ds^2 = dx^2 + dy^2 + dz^2 - c^2 dt^2$, donde c es la velocidad de la luz, que se puede tomar como 1 en un sistema conveniente de unidades. Aquí x, y, z son las coordenadas espaciales, y t , el tiempo. El signo negativo de dt da una expresión formal de la radical y esencial diferencia entre la coordenada del tiempo y las espaciales, es decir, entre el tiempo y el espacio.

En la teoría de la gravitación, el concepto de curvatura del espacio-tiempo juega un papel muy importante. Las ecuaciones fundamentales de la teoría dada por Einstein relacionan de modo inmediato las cantidades que caracterizan la curvatura del espacio-tiempo con las que caracterizan la distribución y movimiento de la materia. Estas ecuaciones son, al mismo tiempo, las del campo gravitatorio; y por tanto, como probó Einstein en colaboración con V. A. Fok, las leyes del movimiento de los cuerpos en un campo gravitatorio se pueden deducir de ellas.

La estructura del espacio-tiempo según la teoría de la relatividad es muy complicada, y el espacio no puede separarse del tiempo ni siquiera en la medida permitida por la teoría especial. Sin embargo, con una cierta aproximación y bajo ciertas condiciones esto sí se puede hacer. El espacio es euclidiano con suficiente aproximación en dominios que son pequeños comparados con la escala cósmica, pero en dominios grandes la diferencia respecto a la geometría euclidiana resulta evidente. Esta diferencia depende de la distribución y movimiento de masas y materia, y alcanza valores apreciables (aunque todavía muy pequeños) en las proximidades de una estrella de gran masa o, en general, cuando las magnitudes de que se trata son de orden cósmico. En algunas hipótesis sobre la estructura del universo como un todo se supone que por término medio la distribución de las masas es aproximadamente uniforme. En una de estas teorías hipotéticas, propuesta por el físico y matemático soviético Fridman, la geometría del espacio en su conjunto coincide con la geometría de Lobachevski.

En la teoría de la relatividad, la geometría abstracta encuentra aplicación no sólo como aparato matemático; las ideas propias de un espacio abstracto proporcionan el medio para una formulación más profunda de los fundamentos de esta teoría. Las posibilidades contempladas en la geometría abstracta se descubren en la realidad, y el pensamiento consigue con ello su triunfo más brillante. La geometría abstracta, nacida del estudio empírico de las relaciones espaciales y formas de los cuerpos, afronta ahora, en calidad de método matemático bien desarrollado, el estudio del espacio real. Tal es el camino general de la ciencia: a partir de lo que la experiencia proporciona directamente, pasa a generalizaciones y abstracciones, volviendo luego otra vez a la experiencia como instrumento para un conocimiento más profundo de la esencia de los fenómenos; y al proporcionar así la explicación de fenómenos conocidos y la predicción de otros nuevos, guía la actividad práctica de los investigadores y a su vez encuentra en ello su propia justificación y la fuente de su futuro desarrollo.

PARTE 6

Capítulo 17

TOPOLOGIA

§1. El objeto de la topología

«La *adyacencia* es una propiedad distintiva de los cuerpos que permite llamarlos *geométricos*, cuando se retiene esta propiedad y se hace abstracción de todas las demás, sean esenciales o accidentales.» Con estas palabras comienza I. N. Lobachevski el primer capítulo de su libro *Nuevos elementos de Geometría*¹.

Explicando mediante un diagrama (figura 1) las palabras citadas, Lobachevski continúa: «Dos cuerpos, *A* y *B*, que se tocan entre sí forman un único cuerpo geométrico *C*... Inversamente, todo cuerpo *C* se puede dividir mediante una sección arbitraria *S* en dos partes, *A* y *B*.»

Los conceptos de adyacencia, vecindad, de lo infinitamente próximo, así como el concepto de disección de un cuerpo, que hasta cierto punto es dual de los primeros, son los que Lobachevski coloca en la base de toda la estructura de la geometría y son también, en esencia, los conceptos fundamentales de la topología

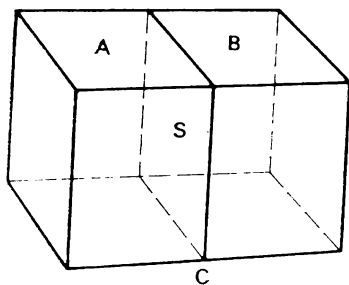


FIG. 1

¹ I. N. Lobachevski, *Collected works*, Gostehizdat, 1949, pág. 168.

en toda la extensión en que ahora entendemos esta disciplina. Por tanto los comentaristas contemporáneos del gran geómetra tienen razón cuando dicen² que «Lobachevski es el autor del primer intento en la historia de las ciencias matemáticas de partir, en la construcción de la geometría, de las propiedades topológicas de los cuerpos. . . Los conceptos de superficie, línea, punto, los define Lobachevski en términos de disecciones y adyacencias de cuerpos». De los siguientes diagramas (Fig. 2), tomados de sus trabajos, se puede obtener una idea de la diversidad de contenido geométrico concreto reflejado en los conceptos de adyacencia y disección de cuerpos tal como los imaginaba Lobachevski.

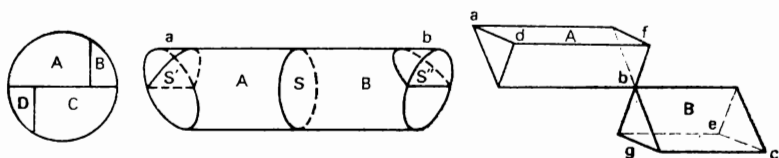


FIG. 2

Toda transformación de una figura que no destruye la adyacencia de las distintas partes de la figura se llama *continua*; si ocurre que no sólo se conservan las adyacencias sino que no se crean otras nuevas, la transformación se llama *topológica*. Por tanto, bajo una transformación topológica de una figura cualquiera, las partes que estaban en contacto siguen en contacto, y las que no lo estaban seguirán sin estarlo; para decirlo brevemente: en una transformación topológica

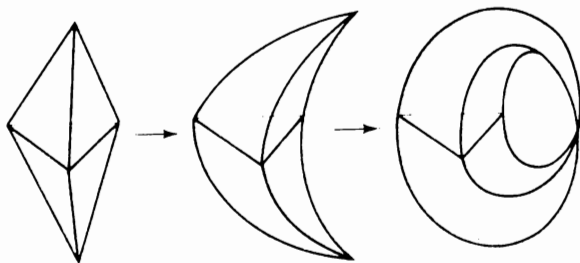


FIG. 3

² Comentarios a *New Elements of Geometry* («Nuevos elementos de Geometría»), loc. cit., pág. 465.

no hay ni roturas ni fusiones. En particular, dos puntos no se pueden unir en uno solo (en este caso aparecerían contactos; figura 3). Por tanto, una transformación topológica de una figura geométrica, considerada como el conjunto de sus puntos, es no sólo continua, sino también uno a uno: cada dos puntos distintos se transforman en otros dos puntos distintos. Así, pues, las transformaciones topológicas son unívocas y continuas en ambas direcciones.

Intuitivamente, podemos representar una transformación topológica de la siguiente manera.

Imaginemos que nuestra figura está hecha de algún material flexible y extensible, por ejemplo, de goma. Podemos entonces someterla a todas las transformaciones continuas posibles, extendiéndola por una parte y contrayéndola por otra, y en conjunto cambiará cada vez de tamaño y de forma. Por ejemplo, si tenemos una gran goma de pelo cuya forma inicial es circular, podemos estirla dándole forma de elipse alargada, podemos darle la forma de un polígono regular o irregular, o incluso de una curva cerrada muy complicada. Algunas de estas posibilidades se ilustran en la figura 4. Pero

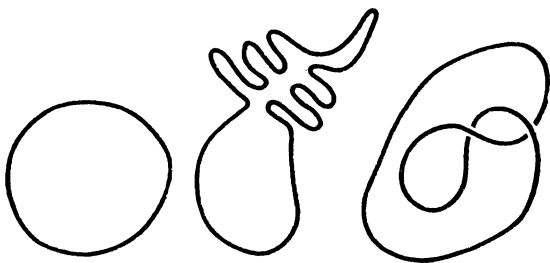


FIG. 4

mediante una transformación topológica nunca podremos convertir un círculo en una figura en forma de ocho (esto requeriría la fusión de dos puntos distintos del círculo; figura 5) o en un intervalo (para lo cual es necesario romper el círculo en un punto cualquiera). El círculo es la curva cerrada más simple, ya que forma un solo lazo, en contraste con la figura en forma de ocho que tiene dos lazos o la curva en forma de trébol que tiene tres (Fig. 5). Las transformaciones topológicas conservan la propiedad del círculo de ser una curva cerrada simple; por ello se dice que esta es una propiedad topológica.

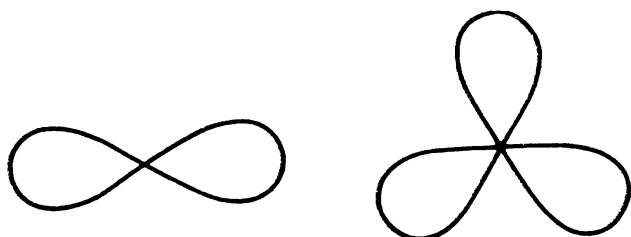


FIG. 5

En una superficie esférica, que podemos suponer hecha de una fina película de goma, podemos inducir también grandes cambios de forma por medio de transformaciones topológicas (Fig. 6). Pero nunca podremos convertirla, mediante una transformación topológica,

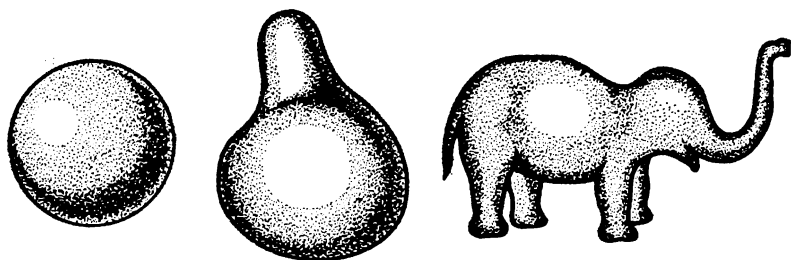


FIG. 6

gica, en un cuadrado o en una superficie en forma de anillo (la superficie de un volante o de un salvavidas) que recibe el nombre de toro (Fig. 7). Para comprobarlo basta ver que la superficie de la esfera tiene dos propiedades que se conservan en las transformaciones topo-

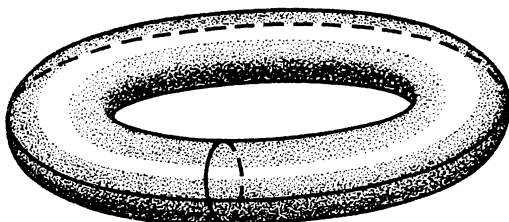


FIG. 7

lógicas. Primero, nuestra superficie es cerrada: no existen aristas en ella (en cambio el cuadrado las tiene); en segundo lugar, toda curva cerrada sobre una superficie esférica es, en expresión de Lobachevski, una disección de ésta; si hacemos un corte a lo largo de una de estas curvas, la superficie queda dividida en dos partes desconexas. El toro no tiene esta propiedad: si lo cortamos a lo largo de un meridiano (Fig. 7), no se divide en dos partes, sino que se convierte en una superficie en forma de tubo doblado (Fig. 8), que se puede convertir

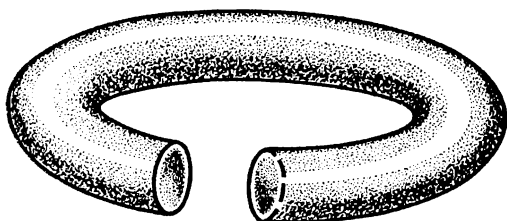


FIG. 8

fácilmente en un tubo cilíndrico mediante una transformación topológica. Así, pues, a diferencia de la esfera, no toda curva cerrada sobre el toro es una disección. Por tanto, la superficie esférica no se puede convertir en un toro mediante una transformación topológica. Decimos que la superficie esférica y el toro son superficies topológicamente distintas, o que son de distinto tipo topológico, o que no son superficies homeomorfas. Por el contrario, una esfera y un elipsoide y, en general, cualquier superficie acotada convexa pertenecen al mismo tipo topológico; es decir, son *homeomorfas*. Esto significa que se puede convertir la una en la otra mediante una transformación topológica.

§2. Superficies

Como ya hemos dicho, toda propiedad de una figura geométrica que se conserva en cualquier transformación topológica recibe el nombre de propiedad topológica. La topología estudia las propiedades topológicas de las figuras, además de las transformaciones topológicas y de las transformaciones continuas de las figuras geométricas.

Acabamos de dar algunos ejemplos de propiedades topológicas. Tales propiedades son: la propiedad de una curva o de una superficie de ser cerrada, la propiedad de ser simple (es decir que forme

un solo lazo), la propiedad de una superficie de que toda curva cerrada sobre ella sea una disección de la superficie (la superficie esférica tiene esta propiedad, pero no así el toro), etc.

El mayor número de curvas cerradas que se pueden trazar sobre una superficie de modo que no formen disecciones, es decir, que la superficie no quede dividida al efectuar un corte a lo largo de todas las curvas, se llama *orden de conexión* de la superficie. Este número suministra importante información sobre la disposición de la superficie. Hemos visto que para la superficie esférica este número es igual a cero (toda curva cerrada sobre esta superficie es una disección). En el toro existen dos curvas cerradas que juntas no forman una disección: un paralelo y un meridiano, por ejemplo (Fig. 7). Sin embargo, es imposible tomar tres curvas de forma que juntas no formen una disección; el orden de conexión del toro es por tanto 2.

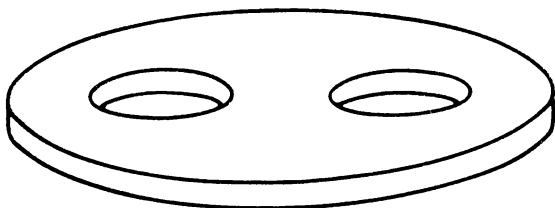


FIG. 9

El orden de conexión de la superficie de la figura 9 es 4, etcétera. De modo general, tomemos una superficie esférica y hagamos $2p$ agujeros circulares en ella (en la figura 10 se ilustra el caso $p = 3$). Dividimos estos agujeros en p pares y pegamos en cada par los extremos de un tubo cilíndrico (un «asa»). Obtenemos así una superficie esférica con p «asas» o, como se dice habitualmente, una superficie normal de género p . El orden de conexión de esta superficie es $2p$.

Todas estas superficies, en expresión de Lobachevski, son «disecciones» del espacio: cada superficie divide el espacio en dos dominios, uno interior y otro exterior, siendo ella misma la frontera común a esos dos dominios. Este hecho está relacionado con otro, a saber, que cada una de las superficies consideradas tiene dos caras (podemos pintar las caras con colores distintos).

Sin embargo, además de éstas, existen también superficies de una sola cara, las cuales no dividen el espacio en dos dominios distintos. La más simple de estas superficies es la llamada «banda de Möbius»,

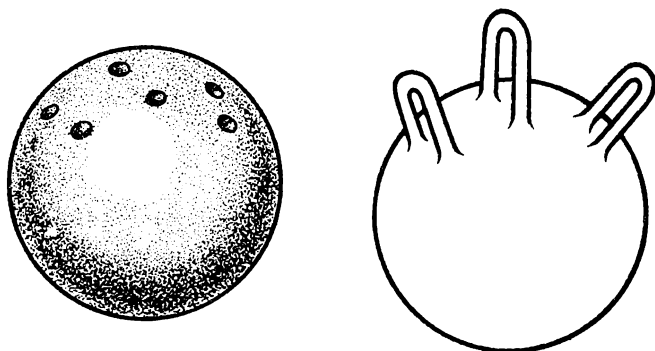


FIG. 10

que se obtiene tomando una tira rectangular de papel $ABCD$ y pegando los bordes opuestos AB y CD de modo que el vértice A coincida con el C y el B con el D . La superficie así obtenida está dibujada en la figura 11; se llama cinta o banda de Möbius. Es fácil comprobar que no tiene dos caras que se puedan pintar de diferentes colores: empezando en el punto E y siguiendo la línea media de la

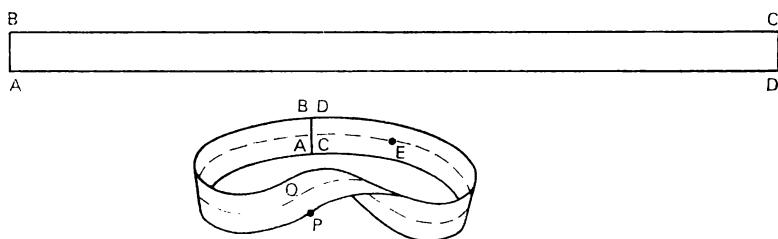


FIG. 11

superficie llegamos de nuevo a E , pero por el otro lado, aunque en el recorrido no hayamos cruzado los bordes. Por otra parte, el borde de la cinta de Möbius consiste en una única línea cerrada.

Surge ahora la siguiente pregunta: ¿existen superficies cerradas de una cara, es decir, superficies de una cara sin bordes? Efectivamente existen, pero tales superficies, por muchas vueltas que les demos

en el espacio tridimensional, tienen siempre autointersecciones. Un ejemplo típico de superficie cerrada de una cara es la dibujada en la figura 12; es lo que se llama «toro de una cara» o botella de Klein.

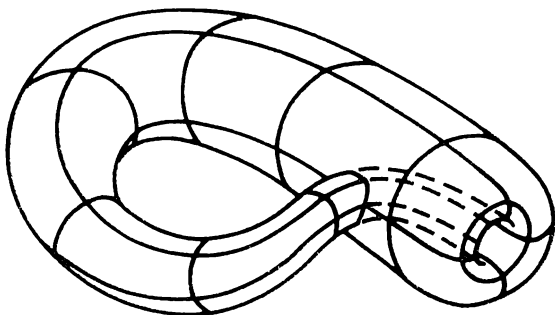


FIG. 12

Si imaginamos, sin preocuparnos de las autointersecciones, dos bandas de Möbius iguales pegadas a lo largo de sus bordes (el borde de una banda de Möbius, como ya hemos dicho, consiste en una única línea cerrada), obtenemos la botella de Klein.

Podemos formular ahora el teorema fundamental de la topología aplicado a las superficies de dos caras: toda superficie de dos caras es homeomorfa a una *superficie normal* de género p , es decir, a una «esfera con p asas»; dos superficies cerradas de dos caras son homeomorfas si, y sólo si, tienen el mismo *género* p (el mismo orden de conexión $2p$), es decir, si son homeomorfas a una esfera con el mismo número de asas p .

Para superficies de una cara existen también «formas normales», similares a las formas normales de las superficies bidimensionales de género p , pero más complicadas de describir. Para ello hay que tomar una esfera, cortar p agujeros circulares y unir a cada uno de ellos una superficie de Möbius pegando el borde de ésta al borde del agujero correspondiente. La dificultad que surge al intentar visualizar este pegado se debe a que su realización física es imposible: en toda materialización de una superficie de una cara en forma de modelo espacial aparecen siempre autointersecciones.

No se debe pensar que las superficies cerradas de una cara pertenecen al dominio de las curiosidades matemáticas y que están desligadas de los problemas serios de la ciencia. Para ver cuán equivo-

cada es esta opinión basta recordar que la geometría proyectiva fue uno de los logros fundamentales del pensamiento geométrico y forma parte hoy de los programas de los cursos universitarios y de las escuelas técnicas. Las fuentes de la geometría proyectiva datan del Renacimiento (Leonardo da Vinci), en relación con las necesidades de la arquitectura, pintura y proyectos técnicos. Los siglos XVI y XVII asistieron al descubrimiento de los primeros teoremas de la geometría proyectiva. Así, naciendo en íntima relación con necesidades prácticas muy definidas, la geometría proyectiva, al desarrollarse por completo, se convirtió en una de las más importantes generalizaciones de la geometría en lo que se refiere a ideas teóricas. En particular, fue en el marco de la geometría proyectiva donde se comprendieron por vez primera las ideas de la geometría no euclidiana de Lobachevski³.

La transición del plano ordinario al plano proyectivo consiste en completar el plano con nuevos elementos abstractos, los llamados puntos impropios o «infinitamente distantes». Al añadir esos puntos, la operación de proyectar un plano sobre otro (por ejemplo, la proyección sobre una pantalla mediante un proyector) se convierte en una transformación uno a uno. El proceso de completar el plano con los puntos impropios, que en la geometría de coordenadas corresponde al paso de coordenadas cartesianas a coordenadas homogéneas, se desarrolla de la siguiente forma. Toda línea recta se completa con un solo punto impropio («en el infinito»), y dos rectas tienen el mismo punto impropio si, y sólo si, son paralelas. Una recta completada con el punto del infinito se convierte en una línea cerrada, y el conjunto de todos los puntos de infinito de todas las posibles rectas forman por definición una línea impropia o línea de infinito.

Puesto que las rectas paralelas tienen en común el punto del infinito, en la representación del proceso de completar el plano con puntos impropios es suficiente considerar las rectas que pasan por un punto arbitrario del plano, por ejemplo, el origen de coordenadas O (Fig. 13). Los puntos impropios de estas rectas agotan ya los puntos impropios de todo el plano proyectivo (puesto que toda recta tiene el mismo punto impropio que la paralela a ella que pasa por O). Obtenemos, por tanto, un modelo del plano proyectivo si lo imaginamos como un círculo de radio «infinitamente» grande con centro en O y si suponemos que todo par de puntos diametralmente opuestos

³ Ver, por ejemplo, capítulo 16, §6, o el libro de P. S. Aleksandrov, *Qué es la geometría no euclidiana*, Moscú, 1951.

A, A' de la circunferencia de este círculo están unidos en el único punto «del infinito» de la recta AA' . La circunferencia de nuestro círculo se convierte entonces en la recta del infinito, pero debemos tener en cuenta que cada par de puntos diametralmente opuestos de esta circunferencia representa en realidad un solo punto. Es evidente, pues, que el plano proyectivo es una superficie cerrada sin bordes.

Si tomamos una curva de segundo orden en el plano proyectivo —una hipérbola, por ejemplo (ver figura 13)— es obvio que, en dicho

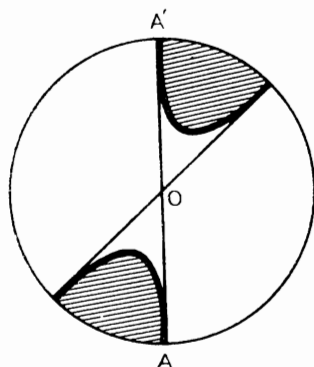


FIG. 13

plano, es una curva cerrada (cortada en dos ramas por la recta del infinito). Teniendo en cuenta que los puntos diametralmente opuestos de la circunferencia de nuestro círculo fundamental son un mismo punto, podemos ver sin dificultad que el interior (sombreado) de la hipérbola en la figura 13 es homeomorfo al interior de un círculo ordinario y que su complemento, la parte no sombreada en la figura, es homeomorfo a una banda de Möbius. Así, pues, desde el punto de vista topológico el plano proyectivo es el resultado de pegar un círculo (en nuestro caso el interior de una hipérbola) con una banda de Möbius a lo largo de sus bordes. De aquí se sigue que el plano

proyectivo, es decir, el objeto básico de estudio de la geometría proyectiva plana, es una superficie cerrada de una cara.

Aparte de su gran valor geométrico intrínseco, el ejemplo del plano proyectivo es también interesante porque pone de relieve una peculiaridad del pensamiento geométrico contemporáneo, en tanto que modelado sobre la base del descubrimiento de Lobachevski. El pensamiento geométrico ha sido siempre abstracto, debido al mismo carácter del concepto de figura geométrica. Ahora se eleva a un nuevo grado de abstracción, que en nuestro caso se hace evidente en el proceso de completar el plano con nuevos elementos abstractos, a saber, los puntos impropios. Ahora bien, incluso estos elementos abstractos tienen una entidad real (cada punto impropio no es más que una abstracción de un haz de rectas paralelas), pero en nuestra discusión aparecen como elementos geométricos distintos que sólo cabe imaginar de modo indirecto como el resultado de la «fusión»

(que físicamente no existe) de puntos diametralmente opuestos de una circunferencia. Construcciones abstractas parecidas tienen gran importancia en el campo de la topología contemporánea, especialmente en la transición de planos a variedades de dos y tres o más dimensiones.

§3. Variedades

Consideremos el siguiente aparato, llamado a veces péndulo plano compuesto (Fig. 14). Consiste en dos barras OA y AB articuladas en A ; el punto O permanece inmóvil, la barra OA gira libremente en un plano fijo alrededor de O y la barra AB gira libremente en el mismo plano alrededor de A . Cada posición de nuestro sistema está completamente determinado por los valores de los ángulos ϕ y ψ que las barras OA y AB forman con una dirección arbitraria fija, por ejemplo la dirección positiva del eje de abscisas. Podemos considerar estos dos ángulos, que varían de 0 a 2π , como las «coordenadas geográficas» de un punto en un toro, contadas a partir del ecuador y de uno de los meridianos del toro, respectivamente (Fig. 15). Podemos decir

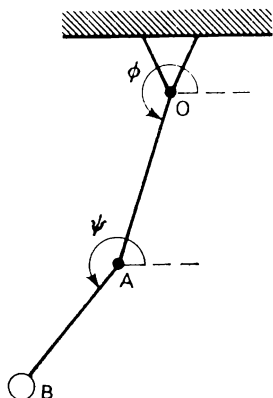


FIG. 14

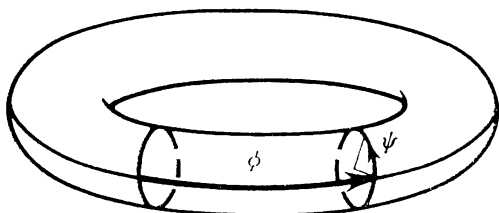


FIG. 15

entonces que todas las posibles posiciones de nuestro sistema mecánico forman una variedad de dos dimensiones, a saber un toro. Si representamos cada uno de los dos ángulos ϕ y ψ por el correspondiente punto en una circunferencia después de fijar sobre ella un punto inicial y una dirección (Fig. 16), podemos decir también que todos los posibles estados de nuestro sistema mecánico quedan completamente caracterizados dando dos puntos, uno en cada

círculo (uno de los círculos se toma como latitud ϕ , y el otro como longitud ψ). En otros términos, exactamente igual que en la geometría analítica se identifica un punto del plano con dos números, que son sus coordenadas, en nuestro caso identificamos un punto de un toro (y por tanto una posición arbitraria del sistema mecánico) con la pareja de sus coordenadas geográficas, es decir, con dos puntos, uno en cada círculo. La esencia de la situación se expresa diciendo que la variedad de todas las posibles posiciones del péndulo plano compuesto es decir, el toro, es el producto topológico de dos círculos.

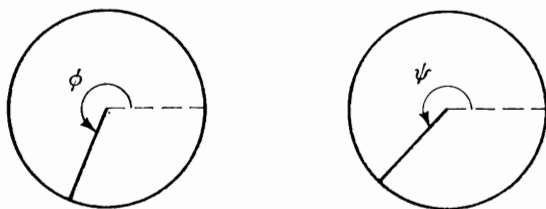


FIG. 16

Modifiquemos ahora el aparato de la siguiente forma. Supongamos que, como antes, consiste en dos barras OA y AB y que OA puede girar libremente en un determinado plano alrededor

de O , pero ahora AB está articulado con OA mediante una rótula esférica en A , de forma que puede girar libremente en el espacio alrededor de A . La posición del nuevo sistema viene dada por tres parámetros, de los cuales el primero es el ángulo ϕ que forma la barra OA con la dirección positiva del eje de abscisas, y los otros dos determinan la dirección de la barra AB en el espacio. Esta dirección se puede especificar, por ejemplo, dando el punto B' (situado sobre la esfera unidad con centro en el origen de coordenadas O) en que el radio OB' paralelo a AB corta a la esfera, o bien dando las dos coordenadas geográficas de B' . Así, pues, la variedad de las posibles posiciones de nuestro sistema articulado es una cierta variedad tridimensional, y el lector podrá fácilmente comprobar que se trata del producto topológico de un círculo y una esfera. Esta variedad es cerrada, es decir, no tiene bordes, y por tanto no puede materializarse en forma de figura en el espacio tridimensional. Si, no obstante, queremos darnos una idea intuitiva de ella, podemos considerar la parte del espacio comprendida entre dos esferas concéntricas. Cada radio que arranca del centro común de las dos esferas las atraviesa en dos puntos. Si consideramos cada pareja de tales puntos como fundidos en uno sólo, obtenemos una variedad tridimensional que es el producto topológico de una esfera por un círculo.

El péndulo articulado se complica todavía más si permitimos que gire no sólo la barra AB sino también la OA en el espacio. El conjunto de todas las posibles posiciones así obtenidas es una variedad cerrada tetradimensional, a saber, el producto de dos esferas.

Vemos, pues, que incluso las consideraciones mecánicas (cinemáticas) más simples conducen ya a variedades topológicas de tres y más dimensiones. De mayor valor aún en la discusión detallada de problemas mecánicos son ciertas variedades (en general multidimensionales) llamadas *espacio de las fases* de sistemas dinámicos. Aquí se tiene en cuenta no sólo las posibles configuraciones del sistema mecánico, sino también las velocidades con que se mueven los distintos puntos que lo constituyen. Limitémonos a uno de los ejemplos más simples. Supongamos que tenemos un punto que se puede mover libremente con velocidad arbitraria sobre una circunferencia. Todo estado de este sistema viene determinado por dos datos: la posición del punto sobre el círculo y la velocidad en el instante dado. La variedad de estados (en este espacio de las fases) del sistema mecánico es, naturalmente, un cilindro infinito (el producto de un círculo por una línea recta).

El número de dimensiones del espacio de las fases aumenta con el número de grados de libertad del sistema mecánico. Muchas propiedades de las características dinámicas de un sistema mecánico se pueden expresar en términos de las propiedades topológicas de su espacio de fases. Por ejemplo, a cada movimiento periódico del sistema corresponde una curva cerrada en su espacio de fases.

El estudio de los espacios de fases de los sistemas dinámicos que aparecen en diversos problemas de mecánica, física y astronomía (mecánica celeste, cosmogonía) atrajo la atención de los matemáticos hacia la topología de variedades multidimensionales. Precisamente en relación con estos problemas, el famoso matemático francés Poincaré comenzó, en la última década del siglo pasado, la construcción sistemática de la topología de variedades multidimensionales, aplicando el llamado método combinatorio, que hasta el momento es uno de los métodos fundamentales en topología.

§4. *El método combinatorio*

Históricamente, el primer teorema de topología es el teorema o fórmula de Euler (que al parecer lo conocía ya incluso Descartes). Consiste en lo siguiente. Tomemos la superficie de un poliedro con-

vexo arbitrario. Denotemos por α_0 , α_1 y α_2 el número de sus vértices, aristas y caras, respectivamente; la siguiente relación se conoce con el nombre de *fórmula de Euler*

$$\alpha_0 - \alpha_1 + \alpha_2 = 2. \quad (1)$$

Este teorema geométrico pertenece a la topología, ya que la fórmula anterior conserva su validez cuando sometemos el poliedro en cuestión a una transformación topológica arbitraria. En una transformación tal las aristas dejarán, en general, de ser rectilíneas, las caras dejarán de ser planas y el poliedro se convertirá en una superficie curva, pero la fórmula (1) entre el número de vértices y aristas y caras (ahora curvas) seguirá siendo válida. El caso más importante es aquel en que todas las caras son triángulos, teniéndose entonces lo que se denomina una *triangulación* (una división de la superficie en triángulos, rectilíneos o curvilíneos). Es fácil reducir el caso general de caras poligonales arbitrarias a este caso: basta dividir estas caras en triángulos (por ejemplo, trazando diagonales desde un vértice cualquiera de una cara dada). Podemos, por tanto, restringir nuestra atención al caso de una triangulación. El método combinatorio en la topología de superficies consiste en sustituir el estudio de una superficie por el de una de sus triangulaciones, interesándonos, claro está, sólo por aquellas propiedades que son independientes de la triangulación elegida; estas propiedades, al ser comunes a todas las triangulaciones, expresan una propiedad de la propia superficie.

La fórmula de Euler conduce a una de tales propiedades, que consideraremos con más detalle. El primer miembro de la fórmula de Euler, es decir, la expresión $\alpha_0 - \alpha_1 + \alpha_2$, donde α_0 es el número de vértices, α_1 el número de aristas y α_2 el número de caras de una triangulación dada, se llama *característica de Euler* de dicha triangulación. El teorema de Euler afirma que para todas las triangulaciones de una superficie homeomorfa a una esfera la característica de Euler es igual a dos. Resulta además que para cualquier superficie (y no sólo para una superficie homeomorfa a la esfera) todas sus triangulaciones tienen la misma característica de Euler.

Es fácil hallar el valor de la característica de Euler para distintas superficies. En primer lugar, para la superficie cilíndrica es igual a cero. Para ello basta comprobar que cuando suprimimos de una triangulación arbitraria de la esfera dos triángulos no adyacentes conservando sus fronteras tenemos evidentemente una triangulación de una superficie homeomorfa a una superficie cilíndrica. Ahora el número

de vértices y aristas es el mismo, pero el número de triángulos ha disminuido en dos, con lo cual el número característico de Euler de la triangulación así obtenida es igual a cero. Tomemos ahora la superficie que se obtiene después de suprimir $2p$ triángulos no adyacentes dos a dos (es decir, que no tienen vértices comunes ni caras comunes) ⁴. Aquí la característica de Euler disminuye en $2p$ unidades. Es fácil ver que la característica de Euler no cambia cuando se adosan tubos cilíndricos («asas») a cada par de agujeros practicados en la superficie de la esfera. Esto se debe a que la característica de los tubos pegados es cero (como ya hemos visto) y sobre el borde del tubo el número de vértices es igual al número de aristas. Así, una superficie cerrada de dos caras de género p tiene como característica de Euler el número $2 - 2p$ (hecho probado en origen por el almirante francés de Jonquières).

Vamos a dar ahora una importante propiedad de las triangulaciones que satisface la condición de invariancia topológica (es decir, toda triangulación de la superficie tiene dicha propiedad si al menos una de ellas la tiene). Se trata de la propiedad de *orientabilidad*. Antes de formularla observemos que todo triángulo se puede orientar, es decir, que se puede especificar una forma definida de recorrerlo. Cada una de las dos posibles orientaciones de un triángulo se determina dando un orden concreto de sucesión de sus vértices ⁵. Supongamos ahora que sobre una superficie nos dan dos triángulos que tienen solamente un lado común (Fig. 17). Dos orientaciones de estos triángulos se llaman compatibles si engendran direcciones opuestas en el lado común. (En el plano o en cualquier otra superficie de dos caras esto significa que, colocada la superficie sobre una de sus caras, los triángulos son recorridos ambos en la misma dirección, esto es, o bien en la dirección de las agujas del reloj, o bien en sentido contrario.) Una triangulación de una superficie cerrada se dice orientable si se puede encontrar una orientación tal de todos los triángulos que las orientaciones de cada dos triángulos adyacentes con un lado común sean compatibles. Se verifica entonces el siguiente resultado: toda triangulación de una superficie de dos caras es orientable, y toda trian-

⁴ Para ello basta hacer la triangulación en cuestión suficientemente «fina». Esto se puede realizar siempre con una subdivisión conveniente de cualquier triangulación.

⁵ Por otra parte es fácil comprobar que dos ordenaciones de los vértices determinan la misma orientación si, y sólo si, se puede transformar la una en la otra mediante una permutación «par». Así, (ABC) , (BCA) , (CAB) determinan una sola orientación del triángulo, y (BAC) , (ACB) , (CBA) otra. (Acerca de permutaciones pares e impares véase, por ejemplo, el capítulo 19, §3).

gulación de una superficie de una cara es no orientable. Por tanto, las superficies de dos caras se llaman también orientables y las de una cara no orientables. Eligiendo una triangulación arbitraria de una banda de Möbius, el lector puede convencerse fácilmente de su no orientabilidad. Para obtener la triangulación más simple del plano proyectivo basta trazar en él tres rectas que no pasen por el mismo punto (Fig. 18). Dividen el plano proyectivo en cuatro triángulos, uno de los cuales está situado en la parte finita del plano mientras que los otros tres están cortados por la recta del infinito en dos partes. En la figura 18 se ha sombreado uno de estos triángulos que se extienden hasta el infinito. De la propia figura se deduce que, si intentamos dar a los cuatro triángulos orientaciones compatibles, estamos condenados al fracaso inevitablemente. En particular, en la orientación dada a los cuatro triángulos de la figura 18 obtenemos que la suma algebraica de sus lados no es cero, como debiera ser si las orientaciones fueran compatibles, sino dos veces la línea AB .

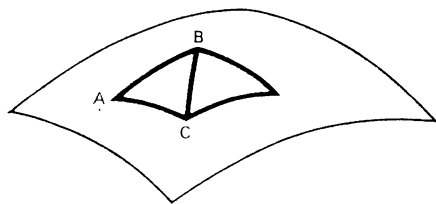


FIG. 17

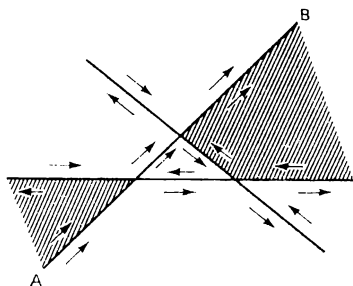


FIG. 18

La característica de Euler y la orientabilidad o no orientabilidad de las superficies cerradas forman, por así decirlo, un sistema completo de invariantes topológicos de las superficies cerradas. El significado de esta afirmación es que dos superficies son homeomorfas si, y sólo si, sus triangulaciones, primero, tienen la misma característica de Euler, y segundo, ambas son orientables o no orientables.

El método combinatorio es aplicable no sólo al estudio de superficies (variedades bidimensionales), sino también a las variedades de un número arbitrario de dimensiones. Pero en el caso de las variedades tridimensionales, el papel de las triangulaciones lo desempeñan

las descomposiciones en tetraedros. Se llaman descomposiciones simpliciales o triangulaciones tridimensionales de la variedad. La característica de Euler de una triangulación tridimensional se define como el número $\alpha_0 - \alpha_1 + \alpha_2 - \alpha_3$ donde $\alpha_i, i = 0, 1, 2, 3$, es el número de elementos i -dimensionales de esta triangulación (es decir, α_0 es el número de vértices, α_1 el número de aristas, α_2 el de fronteras bi-dimensionales y α_3 el número de tetraedros). Para un número de dimensiones $n > 3$ las variedades se dividen en simplex n -dimensionales, esto es, los poliedros n -dimensional convexos más simples análogos a los triángulos ($n = 2$) y a los tetraedros ($n = 3$). Los símlices en que se divide una variedad n -dimensional y sus fronteras forman una triangulación n -dimensional de la variedad. Al igual que antes, podemos hablar de una característica de Euler, interpretándola como la

suma $\sum_{i=0}^n (-1)^i \alpha_i$, donde α_i es el número de elementos i -dimensiona-

les que entran en la triangulación ($i = 0, 1, 2, \dots, n$), y como antes, la característica de Euler tiene el mismo valor para todas las triangulaciones de una variedad n -dimensional dada (y para todas las variedades homeomorfas a ella); es decir, es un invariante topológico. Pero en el estado presente de conocimientos no podemos imaginar un sistema completo de invariantes ni siquiera para las variedades tridimensionales (nos referimos a un sistema análogo al formado por la característica de Euler y la orientabilidad en el caso de las superficies).

El valor del método combinatorio en la topología contemporánea es muy grande. En particular, abre la puerta para la aplicación de ciertos métodos algebraicos en la resolución de problemas topológicos. El lector atento se habrá percatado de esta posibilidad cuando hablábamos anteriormente de la suma algebraica de las fronteras de un triángulo en una triangulación dada del plano proyectivo. Si un triángulo está orientado, esto es, si se ha definido una forma de recorrerlo, es natural tomar como frontera suya la colección de sus lados, cada uno de ellos con la dirección inducida por la orientación.

Consideremos ahora todos los triángulos $T_i^2, i = 1, 2, \dots, \alpha_2$ que entran en una triangulación concreta de una superficie. A cada uno de ellos se le pueden dar dos orientaciones; denotemos el triángulo T_i^2 con una de las dos posibles orientaciones por t_i^2 , y el mismo triángulo con la otra (opuesta) orientación por $-t_i^2$. De la misma forma podemos orientar los elementos unidimensionales (lados) $T_k^1 (k = 1, 2, \dots, \alpha_1)$ que entran en la triangulación dada, es decir, dotarlos de una de las dos posibles direcciones. Denotamos el segmento T_k^1 con una de esas

orientaciones por t_k^1 , y al mismo con la otra orientación por $-t_k^1$. Si los lados del triángulo T_i^2 son T_1^1, T_2^1, T_3^1 , entonces la frontera del triángulo orientado t_i^2 es el conjunto de esos mismos lados pero tomados con una dirección definida, es decir, la frontera consiste en los segmentos dirigidos $\epsilon_1 t_1^1, \epsilon_2 t_2^1, \epsilon_3 t_3^1$; aquí $\epsilon_i = 1$, si esta dirección para el lado T_i^1 es la de t_i^1 , y $\epsilon_i = -1$ en el caso opuesto. La frontera de t_i^2 se denota por Δt_i^2 . Como hemos visto, $\Delta t_i^2 = \sum \epsilon_k t_k^1$ pudiendo suponerse que esta suma se extiende a todos los lados de nuestra triangulación si consideramos los coeficientes ϵ_k correspondientes a segmentos que no están en la frontera de t_i^2 iguales a cero.

Parece ahora natural considerar sumas más generales de la forma $x^1 = \sum a_k t_k^1$ extendidas a todos los lados de nuestra triangulación⁶. El significado geométrico de tales sumas es muy simple: cada sumando es un cierto segmento que entra en la triangulación afectado de una determinada dirección y de un coeficiente concreto (una «multiplicidad» concreta).

La suma algebraica total así descrita representa un camino compuesto de segmentos en el cual cada segmento entra tantas veces como su coeficiente indica. Por ejemplo, si comenzamos por recorrer el polígono $ABCDEF$ (figura 19) en la dirección de la flecha, con-

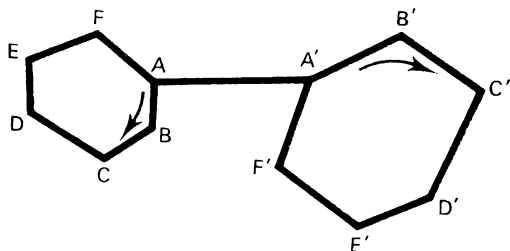


FIG. 19

continuamos por AA' hasta el polígono $A'B'C'D'E'F'$, recorreremos éste en la dirección indicada, volvemos por $A'A$ y de nuevo recorremos $ABCDEF$ como antes, obtenemos una suma en la cual los segmentos AB, BC, CD, DE, EF, FA entran con coeficiente 2; los segmentos $A'B', B'C', C'D', D'E', E'F', F'A'$, con coeficiente 1, y el segmento AA' no aparece en absoluto (tiene coeficiente cero porque está recorrido dos veces en direcciones opuestas).

Las sumas de la forma $x^1 = \sum a_k t_k^1$ se llaman *cadena unidimensional* de la triangulación dada. Desde el punto de vista algebraico

⁶ Los coeficientes a_k se suponen enteros.

representan formas lineales (polinomios homogéneos de primer grado); se pueden sumar y restar y también multiplicar por un número entero arbitrario de acuerdo con las reglas usuales del álgebra. Entre las cadenas unidimensionales tienen especial importancia los ciclos unidimensionales. Geométricamente corresponden a caminos cerrados (precisamente el camino de la figura 19 era de este tipo).

Para una definición puramente algebraica de ciclo convengamos en que de los dos vértices del segmento $\overrightarrow{AB}$, el punto final B entra en la frontera de AB con signo más (con coeficiente $+1$) y el punto inicial A con signo menos (con coeficiente -1). Entonces la *frontera* del segmento $\overrightarrow{AB}$ se puede escribir en la forma $\Delta(AB) = B - A$.

Si aceptamos este convenio, observamos inmediatamente que la suma de las fronteras de los segmentos $\overrightarrow{AB}, \overrightarrow{BC}, \overrightarrow{CD}, \dots, \overrightarrow{FA}$, que forman un camino cerrado (en el sentido usual de la palabra) es cero. Esto conduce de forma natural a la definición general de *ciclo* unidimensional como una cadena unidimensional $z^1 = \sum_k a_k t_k^1$ para la cual la suma de las fronteras de sus términos es cero, o sea, $\sum_k a_k \Delta t_k^1 = 0$. Es fácil verificar que la suma de dos ciclos es también un ciclo. Si multiplicamos un ciclo como una expresión algebraica por un número entero, obtenemos de nuevo un ciclo. Esto permite hablar de combinaciones lineales de ciclos $z_1^1, z_2^1, \dots, z_s^1$, es decir, de ciclos de la forma $z = \sum_{v=1}^s c_v z_v^1$, donde los c_v son enteros.

Por analogía con el concepto de cadena unidimensional de una triangulación dada, podemos hablar también de cadenas bidimensionales de esta triangulación, es decir, de expresiones de la forma $x^2 = \sum_i a_i t_i^2$, donde los t_i^2 son triángulos orientados de la triangulación. Puesto que la frontera de cada triángulo orientado es un ciclo unidimensional, la cadena $\sum_i a_i \Delta t_i^2$ es también un ciclo. Este ciclo se toma como frontera Δx^2 de la cadena $x^2 = \sum_i a_i t_i^2$.

El concepto de *frontera de una cadena* permite formular el concepto de homología: un ciclo unidimensional z^1 de una triangulación dada se dice *homólogo a cero* en esta triangulación si es la frontera de alguna cadena bidimensional de esta triangulación. En toda triangulación de una superficie cerrada convexa y, más generalmente, de toda superficie homeomorfa a una esfera, todo ciclo unidimensional es homólogo a cero; geométricamente esto es perfectamente claro: todo polígono cerrado sobre una superficie convexa es la frontera de algún trozo de la superficie. Sin embargo, esto no es así

en el toro: ni los meridianos del toro, ni su ecuador, son fronteras de ningún trozo de superficie. Si tomamos una triangulación arbitraria del toro, podemos encontrar ciclos similares a un meridiano o al ecuador del toro, y estos ciclos no son homólogos a cero.

En la triangulación del plano proyectivo que hemos construido observamos un fenómeno enteramente nuevo. Si consideramos una línea recta —por ejemplo AB (ver Fig. 18)— como un ciclo de esa triangulación, se observa que no es homólogo a cero en ella. Sin embargo, la misma línea tomada con coeficiente 2 resulta ser un ciclo homólogo a cero. Así, la introducción de coeficientes distintos de $+1$, que en principio podría parecer formal e innecesaria, permite discernir importantes propiedades geométricas de las superficies y, más generalmente, de las variedades. En el caso dado se trata de la llamada propiedad de torsión, que consiste en la existencia de ciclos que no son homólogos a cero en la variedad dada (no limitan ningún trozo de superficie), pero que sí lo son cuando se les afecta de ciertos coeficientes enteros.

En relación con lo que hemos dicho, introduzcamos el concepto, extremadamente importante, de *independencia homológica* de ciclos. Los ciclos $z_1, \dots, z_n$ se dice que son homológicamente independientes en una triangulación si no existe ninguna combinación lineal $\sum c_i z_i$ de ellos en la cual al menos uno de los coeficientes c_i sea distinto de cero y que sea homóloga a cero en la triangulación dada. Podemos tomar como ejemplo de ciclos homológicamente independientes sobre el toro un meridiano arbitrario y el ecuador considerados como ciclos de alguna triangulación de esta figura.

Los conceptos fundamentales de toda la topología combinatoria (los conceptos de frontera, ciclo, homología) han sido definidos para formaciones unidimensionales, pero se pueden extender al pie de la letra a un número arbitrario de dimensiones. Por ejemplo, una cadena bidimensional $z^2 = \sum a_i t_i^2$ se llama ciclo si su frontera $\Delta z^2 = \sum a_i \Delta t_i^2$ es igual a cero. Una cadena tridimensional es una expresión de la forma $x^3 = \sum a_h t_h^3$, donde t_h^3 son símplexes tridimensionales orientados (tetraedros).

Como en el caso del triángulo, la orientación del simplex tridimensional (tetraedro) viene dada mediante una ordenación concreta de sus vértices, donde dos órdenes determinan la misma orientación si se puede transformar el uno en el otro mediante una permutación par. La frontera de un simplex tridimensional orientado $t^3 = (ABCD)$ es la cadena bidimensional (ciclo) $\Delta t^3 = (BCD) - (ACD) + (ABD) -$

—(ABC) (Fig. 20). La frontera de una cadena tridimensional se define como la suma de las fronteras de sus símlices tomados con los mismos coeficientes con que entran en la cadena dada. El lector podrá comprobar fácilmente que la frontera de una cadena tridimensional es un ciclo bidimensional (es suficiente probarlo para un único simplex tridimensional). Decimos que un ciclo bidimensional es homólogo a cero en una cierta variedad si es la frontera de alguna cadena tridimensional de esta variedad, y así sucesivamente. Observemos que de la definición de triangulación orientable y no orientable⁷ dada previamente se sigue que en toda triangulación orientable existen ciclos (bidimensionales en el caso de superficies) diferentes de cero, y que en las no orientables no existen tales ciclos; este resultado se puede generalizar inmediatamente a un número arbitrario de dimensiones.

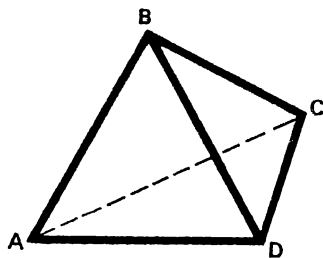


FIG. 20

Los conceptos introducidos permiten definir el orden de conexión unidimensional, bidimensional, etc., de variedades de un número cualquiera de dimensiones. El máximo número de ciclos unidimensionales, bidimensionales, etc., homológicamente independientes que existen en una triangulación arbitraria de una variedad dada no depende de la elección de la triangulación en dicha variedad y recibe el nombre de *orden de conexión* o número de Betti de la variedad (de la correspondiente dimensión).

El número de Betti unidimensional de una superficie orientable cerrada de género p es $2p$ (es decir, el orden de conexión de la superficie, tal y como lo definimos en el §2). El número de Betti unidimensional del plano proyectivo es igual a cero. (Aquí, todo ciclo que no se extiende hasta el infinito limita una parte del plano, esto es, es homólogo a cero; pero un ciclo que se extiende hasta el infinito —por ejemplo, una línea proyectiva— resulta ser homólogo a cero si se toma dos veces.) El número de Betti bidimensional de toda superficie no orientable es cero (no hay ningún ciclo bidimensional distinto de cero sobre la superficie).

⁷ Esta definición, dada anteriormente para triangulaciones de superficies, se puede extender al caso de triangulaciones de variedades de un número cualquiera de dimensiones.

El número de Betti bidimensional de toda superficie orientable es igual a uno, pues, si orientamos todos los triángulos de una triangulación arbitraria de una superficie orientable de manera conveniente, obtenemos entonces un ciclo (el llamado ciclo fundamental de la superficie). No es difícil observar que todo ciclo bidimensional se obtiene a partir de un ciclo fundamental multiplicando éste por un número entero. Estos resultados se pueden generalizar inmediatamente a variedades n -dimensionales. Hacemos notar que el número cero-dimensional de Betti de una variedad conexa (que no está dividida en partes) se toma como 1.

Los números de Betti de las distintas dimensiones se relacionan con la característica de Euler por una notable fórmula que fue probada por Poincaré y que generaliza el teorema de Euler. Esta fórmula conocida por el nombre de *fórmula de Euler-Poincaré*, tiene la siguiente forma:

$$\sum_{r=0}^n (-1)^r \alpha_r = \sum_{r=0}^n (-1)^r p_r.$$

En el primer miembro tenemos la característica de Euler de una triangulación arbitraria de la variedad dada, y los números p_r de la derecha son los números de Betti de las distintas dimensiones r de la variedad.

En la particular para superficies orientables tenemos, como acabamos de ver, $p_0 = p_2 = 1$, $p_1 = 2p$, donde p es el género de la superficie. La fórmula nos da entonces el teorema de Euler para superficies orientables

$$\alpha_0 - \alpha_1 + \alpha_2 = 2 - 2p.$$

§5. Campos vectoriales

Consideremos la ecuación diferencial más sencilla

$$\frac{dy}{dx} = F(x, y), \quad (2)$$

en un dominio plano G . Su significado geométrico es que en todo punto (x, y) de G está definida una dirección cuya pendiente es igual a $F(x, y)$, donde $F(x, y)$ es una función continua del punto (x, y) . Decimos entonces que en G se ha dado un campo continuo de direcciones; podemos convertirlo fácilmente en un campo continuo de

vectores, tomando, por ejemplo, un vector de longitud unidad en cada una de las direcciones. La tarea de integrar la ecuación diferencial (2) consiste en dividir, si es posible, el dominio plano dado en curvas que no se corten dos a dos (las «curvas integrales» de la ecuación), de forma que en cada punto del dominio la dirección dada sea la dirección de la tangente a la única curva integral que pasa por él.

Consideremos, por ejemplo, la ecuación

$$\frac{dy}{dx} = \frac{y}{x}.$$

La dirección correspondiente a cada punto del plano $M(x, y)$ es evidentemente la del rayo $\overrightarrow{OM}$ (donde O es el origen de coordenadas). Las curvas integrales son las líneas rectas que pasan por O . Por cada punto del plano pasa una única integral. En cuanto al origen de coordenadas, es un punto singular (un «nodo») de la ecuación diferencial por el que pasan todas las curvas integrales.

Si tomamos la ecuación diferencial

$$\frac{dy}{dx} = -\frac{x}{y},$$

vemos que asocia a cada punto $M(x, y)$ distinto de O la dirección perpendicular a $\overrightarrow{OM}$. En este caso las curvas integrales son circunferencias con centro en O , que de nuevo es un punto singular de nuestra ecuación diferencial, pero de naturaleza completamente diferente. No es un «nodo», sino un punto de los llamados «centro». Existen también otros tipos de puntos singulares (ver capítulo 5, §6), algunos de los cuales se ilustra en las figuras 21 y 22. La ecuación diferencial $dy/dx = y/x$ no tiene curvas integrales cerradas. Por el contrario, la ecuación $dy/dx = -x/y$ sólo tiene curvas integrales cerradas. También hay curvas integrales que rodean en forma de espiral un punto singular; éste recibe entonces el nombre de foco.

De gran importancia en diversas aplicaciones es el caso del llamado ciclo límite, esto es, una curva integral cerrada alrededor de la cual se enrollan otras curvas en espiral. Existen además otros muchos casos de posición mutua de curvas integrales, así como de posiciones respecto a los puntos singulares. Todos los problemas relativos a la forma y posición de las curvas integrales de ecuaciones diferenciales, y también al número, carácter y posición relativa de los

puntos singulares, pertenecen a la teoría cualitativa de ecuaciones diferenciales. Como el mismo nombre indica, la teoría cualitativa de ecuaciones diferenciales deja de lado la integración directa de las ecuaciones «en forma finita», así como también los métodos de aproximación o integración numérica. El objeto básico de esta teoría es, en esencia, la topología del campo de direcciones y el sistema de curvas integrales de la ecuación diferencial dada.

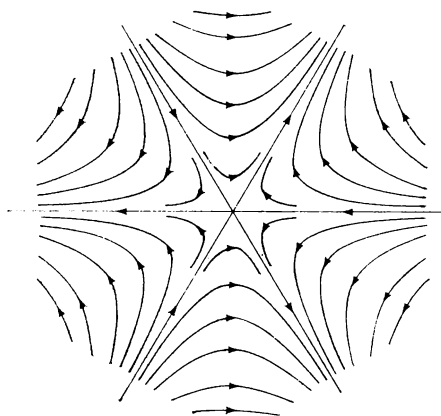


FIG. 21

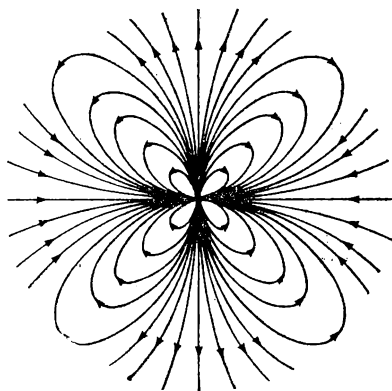


FIG. 22

El enfoque cualitativo de la teoría de ecuaciones diferenciales, incluyendo problemas tales como la existencia de curvas integrales cerradas, y en particular todos los problemas relacionados con la existencia, número y casos de ciclos límite, vino impuesto en principio por problemas de mecánica, física y tecnología. Estos problemas surgieron por vez primera en relación con los estudios de Poincaré en mecánica celeste y cosmogonía, los cuales, como ya hemos dicho, fueron también el motivo para las investigaciones topológicas del geómetra francés. El campo de los problemas topológicos en la teoría de ecuaciones diferenciales ha sido objeto principal de los notables trabajos de los profesores soviéticos en la teoría de oscilaciones y radiotecnología; nos referimos a la escuela de L. I. Mandelstam, así como a la de A. A. Andronov, que se ha desarrollado a partir de aquella y que constituye uno de los principales centros de investigación en la teoría cualitativa y esencialmente topológica de ecua-

ciones diferenciales. Otro centro de investigación por métodos esencialmente topológicos es la escuela de V. V. Stepanov y V. V. Nemyski en la Universidad de Moscú. En las universidades de Leningrado, Sverdlovsk y Kazan, los métodos topológicos juegan también un papel dominante en los trabajos matemáticos sobre la teoría cualitativa de ecuaciones diferenciales.

La teoría de ecuaciones diferenciales conduce al estudio de campos vectoriales, no sólo en el plano, sino también en variedades multidimensionales; incluso los sistemas más simples de ecuaciones diferenciales se interpretan geoméricamente como campos de direcciones en espacios euclídeos multidimensionales. La introducción de las llamadas integrales primeras de una ecuación significa una selección, de entre el conjunto de todas las curvas integrales, de aquellas que están en una cierta variedad definida por la integral primera dada. Todo sistema dinámico (en el sentido clásico de la palabra) da lugar, en general, a la variedad multidimensional de sus posibles estados (ver §3) y a un sistema de ecuaciones diferenciales cuyas curvas integrales, que llenan el espacio de las fases dado, representan posibles movimientos del sistema. Cada uno de estos posibles movimientos está determinado por un cierto conjunto de condiciones iniciales. Por tanto, en este caso, es objeto fundamental de estudio el campo de direcciones y el sistema de trayectorias sobre la variedad dada. Un sinnúmero de aplicaciones, especialmente en años recientes, han aconsejado un amplio desarrollo de la teoría cualitativa de ecuaciones diferenciales, con el consiguiente desarrollo de la topología como base de esta teoría. Precisamente campos como la mecánica física y la astronomía han sido la causa del rápido crecimiento de la topología, que tan importante parte tiene en el desarrollo de las matemáticas de esta segunda mitad del siglo. El lector que desee familiarizarse con los problemas topológicos de la teoría de ecuaciones diferenciales y sus aspectos físicos y tecnológicos concretos debe consultar el conocido libro de A. A. Andronov y S. E. Haikin *Theory of oscillations*, publicada en lengua inglesa por Princeton University Press, Princeton, N. J., 1949.

Como ejemplo de problema concreto en la teoría de campos vectoriales sobre variedades multidimensionales consideremos el problema del número algebraico de puntos singulares de un campo vectorial.

Supongamos una variedad lisa. Para mayor simplicidad supongamos que es una superficie lisa y cerrada. Supongamos además que en todo punto de la variedad está definido un vector tangente a ella

que depende continuamente del punto, tanto en longitud como en dirección. Los puntos singulares de este campo vectorial son los puntos de la variedad en los que el vector asociado a ellos es nulo, es decir, aquellos puntos en los que no está definida una dirección. Supondremos que cada uno de estos puntos es aislado. En el caso de una variedad cerrada esto significa que el número de dichos puntos es finito. (De no ser así, parte de esos puntos estarían agrupados en las proximidades de un cierto punto límite, que por continuidad del campo sería también un punto singular, pero no aislado.)

Definamos ahora el índice de un punto singular aislado, concepto que en cierto sentido es similar al de multiplicidad de una raíz de una ecuación algebraica. Para definir el índice redeamos el punto singular con una curva arbitraria cerrada C que «lo aisle» (en el plano sería simplemente un círculo), esto es, una curva que no pasa por ningún otro punto singular y cuyo interior contiene sólo el punto singular en cuestión. En todos los puntos de C la dirección del campo está unívocamente determinada. Suponemos, por sencillez, que hay un entorno *plano* del punto singular que contiene a C (en el caso general podemos transformar el entorno que nos interese, junto con el campo dado en él, en un plano). Si damos una vuelta completa alrededor de la curva en la dirección positiva, el ángulo que forma la dirección del campo con una dirección fija arbitraria vuelve a su valor inicial incrementado en un cierto sumando de la forma $2k\pi$, siendo k un número entero bien definido. Este número recibe el nombre de *índice del punto singular* o *número de vueltas* del campo a lo largo de C . Señalemos que no depende de la elección de la curva que aisle el punto singular. En las figuras 21 y 22 los puntos singulares tienen índices $+3$ y -2 , respectivamente. De forma similar, pero más complicada, podemos también definir el *índice* de puntos singulares de un campo de vectores (direcciones) definidos en una variedad n -dimensional para $n > 2$. El matemático alemán Hopf demostró en 1926 el siguiente teorema: dado un campo continuo de vectores (sobre una variedad) que tiene solamente un número finito de puntos singulares, la suma de sus índices, o como también se dice, el número algebraico de esos puntos singulares, no depende del campo y es siempre igual a la característica de Euler de la variedad.

Del teorema de Hopf se sigue que los campos vectoriales sin singularidades sólo son posibles sobre variedades cuya característica es igual a cero; resulta que sobre tales variedades siempre se puede construir un campo vectorial sin singularidades. Así, entre todas las superficies cerradas, solamente sobre el toro y sobre la  botella de

Klein (llamado toro de una cara) se pueden construir campos de vectores sin singularidades.

Estrechamente relacionada con la teoría de campos vectoriales están la teoría de aplicaciones continuas de variedades en sí mismas y, en particular, los resultados relativos a la existencia de puntos fijos para tales aplicaciones. Se dice que el punto x es un punto fijo para la aplicación f si su imagen coincide con él mismo, es decir, si

$$f(x) = x.$$

Para comprender el carácter de esta conexión, consideremos el caso más simple, esto es, el caso de una aplicación continua f de un círculo K en sí mismo. Uniendo cada punto x con su imagen $f(x)$ obtenemos un vector $\vec{u_x} = \overrightarrow{x, f(x)}$. Este vector es el vector nulo si, y sólo si, $f(x) = x$, es decir, si x es un punto fijo de la aplicación dada. Mostremos que efectivamente existe un punto fijo. Para ello supongamos lo contrario y determinemos el número de vueltas de nuestro campo vectorial a lo largo de la circunferencia C del círculo K .

Es evidente que el número de vueltas de nuestro campo a lo largo de C no puede variar, bajo una transformación continua, más que de manera continua. Pero como es entero debe permanecer constante, de donde se sigue que el número de vueltas del campo a lo largo de C es igual a 1. Efectivamente, puesto que todo punto de K se aplica en el interior del círculo, el vector $\vec{u_x}$ de un punto x sobre la circunferencia C (que, de acuerdo con lo supuesto, no es el vector nulo) está dirigido hacia el interior del círculo y forma por tanto un ángulo agudo con el radio Ox , que podemos considerar como un vector dirigido hacia el centro O .

Sometemos ahora a las direcciones de todos los vectores $\vec{u_x}$ correspondientes a puntos x de C a una transformación continua. Esta transformación consiste en girar dichos vectores un ángulo tal que queden dirigidos hacia el centro O . Como ya hemos dicho, el número de vueltas del campo a lo largo de C no cambia en este proceso. Pero, como resultado de esta transformación, nuestro campo original sobre C se convierte en un campo de vectores radiales, que evidentemente tiene como número de vueltas la unidad. Por tanto, el campo inicial también tiene un número de vueltas igual a 1 a lo largo de la circunferencia C .

En virtud de la continuidad del campo original de vectores, su número de vueltas a lo largo de dos circunferencias concéntricas en O

y de radios ligeramente diferentes es el mismo ⁸. Por tanto, todas las circunferencias con centro en O dentro del círculo K dan el mismo número de vueltas, esto es, uno. Pero como hemos supuesto que el vector u_x está definido y es diferente de cero para todos los puntos del círculo y entre ellos el centro O , el número de vueltas del campo a lo largo de circunferencias de radios suficientemente pequeños con centro en O debe ser cero. Llegamos así a una contradicción, y hemos probado por tanto que una aplicación continua de un círculo en sí mismo tiene al menos un punto fijo. Este teorema es un caso especial del importantísimo teorema de Brouwer que afirma que toda aplicación continua de una esfera n -dimensional en sí misma tiene al menos un punto fijo.

En estos últimos años el problema de la existencia de puntos fijos para aplicaciones de uno u otro tipo ha sido estudiado en detalle y forma una parte esencial de la topología de variedades.

§6. *El desarrollo de la topología*

La topología de superficies cerradas es el único campo de la topología que estaba ya más o menos terminado a finales del siglo pasado. La construcción de esta teoría estuvo relacionada con el desarrollo, a lo largo del siglo XIX, de la teoría de funciones de una variable compleja. Esta teoría, que es uno de los fenómenos más importantes en la historia de las matemáticas del siglo pasado, ha sido construida por distintos métodos. Uno de los más fructíferos, si atendemos a la comprensión de los fenómenos a estudiar, fue el método geométrico de Riemann. El método de Riemann, que mostró muy convincentemente que en la teoría general de funciones de una variable compleja es imposible limitarse a funciones unívocas solamente, condujo a la construcción de las llamadas superficies de Riemann. En el caso más simple de funciones algebraicas de una variable compleja, estas superficies resultan ser siempre superficies cerradas orientables. El estudio de sus propiedades topológicas es, en cierto sentido, equivalente al estudio de la propia función algebraica. El desarrollo ulterior de la idea de Riemann se debe a Poincaré, Klein y sus discípulos, y conduce al descubrimiento de inesperadas y profundas conexiones entre

⁸ En virtud de la hipótesis de que el campo en cuestión, que está definido y es diferente de cero para todo valor de la variable, no tiene puntos fijos para f , podemos también hablar de su número de vueltas a lo largo de una curva arbitraria dentro de K .

la teoría de funciones, la topología de superficies cerradas y las geometrías no euclidianas, a saber, la teoría del grupo de movimientos en un plano de Lobachevski⁹. De esta forma, la topología demostró, desde su aparición, estar orgánicamente relacionada con todo un grupo de problemas fundamentales ligados a muy diversos campos de las matemáticas.

En desarrollos posteriores de este círculo de problemas resultó que la topología de superficies, por sí sola, era insuficiente, y que era necesario estudiar la solución de problemas en topología n -dimensional. El primero de ellos fue el de la invariancia topológica de la dimensión de un espacio. Este problema consiste en probar la imposibilidad de una aplicación topológica entre un espacio euclídeo n -dimensional y otro m -dimensional para $n \neq m$. El problema, muy difícil, fue resuelto en 1911 por Brouwer¹⁰. En relación con su solución se descubrieron nuevos métodos topológicos que condujeron a una rápida construcción de los principios de la teoría de aplicaciones continuas y campos vectoriales sobre variedades multidimensionales. Todos estos estudios resultaron implicar los conceptos fundamentales de la llamada topología conjuntista, que nace sobre la base de la teoría general de conjuntos construida por Cantor en el último cuarto del siglo XIX.

El objeto de estudio de la topología conjuntista, es decir, la clase de figuras geométricas que se consideran, es extremadamente amplio y comprende, si no todos los conjuntos de los espacios euclídeos, al menos todos los conjuntos cerrados. Matemáticos de muchos países colaboraron en el rápido desarrollo de la topología en esta nueva dirección, pero sobre todas debemos mencionar la escuela topológica polaca.

En los trabajos de los topólogos soviéticos se desarrolló el estudio de la topología conjuntista en una dirección esencialmente nueva; en particular, el notable matemático soviético P. S. Uryson (1898-1924), fallecido prematuramente, desarrolló la teoría general de la dimensión, que puso los cimientos para una clasificación de conjuntos de puntos muy generales mediante el criterio fundamental de su número

⁹ En relación con esto se puede consultar el libro de A. I. Markusevic *Teoría de funciones analíticas*, Gostehizdat, 1950.

¹⁰ En realidad, para el desarrollo de la teoría de funciones de una variable compleja, fue necesario resolver un problema aún más difícil, a saber, el de probar que la imagen topológica de un dominio n -dimensional contenido en un espacio n -dimensional es también un dominio. Este problema fue también resuelto por Brouwer.

de dimensiones. Esta clasificación resultó ser extremadamente fructífera e insinuó puntos de vista enteramente nuevos en el estudio de las formas geométricas más generales¹¹. Las ideas de Uryson, tal como las desarrolló en su teoría de la dimensión, sirvieron de punto de apoyo para el notable trabajo de L. A. Liusternik (juntamente con L. G. Snirelman) sobre el cálculo de variaciones.

Estos trabajos contienen, entre otros resultados, una solución completa y positiva del famoso problema de Poincaré sobre la existencia de tres líneas geodésicas cerradas sin puntos múltiples sobre toda superficie homeomorfa a la esfera.

Por otra parte, P. S. Aleksandrov, sobre la base de la teoría de la dimensión, trasplantó los métodos algebraicos de la topología combinatoria al campo de la teoría de conjuntos, lo cual dio lugar a su vez a nuevas direcciones en las investigaciones topológicas en las que matemáticos soviéticos, incluidas las generaciones más jóvenes, desempeñan un papel predominante¹².

En cuanto a la propia topología combinatoria, después de los estudios de Poincaré y Brouwer, hacia 1915, se inicia una serie de investigaciones a cargo de un grupo de topólogos norteamericanos, entre ellos Veblen, Birkoff, Alexander y Lefschetz, logrando resultados muy notables. Por ejemplo, Alexander probó la invariancia topológica de los números de Betti y también un importante teorema de dualidad que sirvió de punto de partida a las posteriores investigaciones de L. S. Pontriagin; Lefschetz dio una fórmula para el número algebraico de puntos fijos bajo aplicaciones continuas de variedades, colocando así los fundamentos de la teoría general algebraica de aplicaciones continuas, desarrollada posteriormente por Hopf; a Birkoff le debe nuestra ciencia un avance esencial en la teoría de

¹¹ La definición inductiva de dimensión de conjuntos propuesta por Uryson se puede considerar como el desarrollo más completo de la idea (propuesta por Lobachevski) de la disección como la operación geométrica fundamental. *Grosso modo* viene a ser lo siguiente. Un conjunto tiene dimensión cero si se puede representar en forma de una suma de partes arbitrariamente pequeñas de las cuales cada dos no están en contacto. Un conjunto tiene dimensión n si se puede «diseccionar» por conjuntos de dimensión $n - 1$ en partes arbitrariamente pequeñas, cada dos de las cuales no están en contacto, y si además esto no se puede realizar con conjuntos de dimensión menor que $n - 1$. (En el §7 se dará una definición precisa de contacto tal como se entiende en la topología contemporánea.)

¹² Mencionemos aquí la llamada teoría homológica de la dimensión, de P. S. Aleksandrov, las notables construcciones de L. N. Pontriagin y los posteriores desarrollos de dicha teoría en los trabajos de M. S. Bokstein, V. G. Boltjanski, y sobre todo K. A. Sitnikov. La ley de dualidad de L. S. Pontriagin se discutirá más adelante.

sistemas dinámicos en su aspecto puramente topológico, en su aspecto métrico, etc. Por otro lado, Hopf desarrolló profundamente la topología de variedades y sus aplicaciones continuas, probando, junto a otros muchos resultados, la existencia de un número infinito de aplicaciones continuas de una esfera tridimensional en otra bidimensional, aplicaciones que son esencialmente distintas en el sentido de que no se puede transformar una en otra mediante un cambio continuo. Hopf se convirtió así en el fundador de una nueva dirección, la llamada topología homotópica (homotopía). Recientemente, la topología homotópica, así como la topología combinatoria, han recibido un nuevo y poderoso impulso a raíz de los trabajos de la nueva escuela topológica francesa (Leray, Serre y otros).

Como ya hemos dicho anteriormente, las investigaciones de Uryson marcaron el comienzo de las actividades de los matemáticos soviéticos en el campo de la topología. Al principio, abordaron en sus trabajos la teoría de la topología de conjuntos, pero ya al final de los años veinte sus intereses se dirigieron también hacia la topología combinatoria. Esto se produjo de una manera bastante original: a través de la aplicación de métodos combinatorios al estudio de conjuntos cerrados, es decir, a objetos de una naturaleza muy general. Sobre esta base surgió uno de los descubrimientos geométricos más notables del presente siglo: el enunciado y demostración por L. S. Pontriagin de su ley general de dualidad, que establece profundas —y en cierto sentido exhaustivas— relaciones entre la estructura topológica de un conjunto cerrado en un espacio euclídeo n -dimensional y su complementario. En conexión con su ley general de dualidad, Pontriagin construyó una teoría general de caracteres de los grupos conmutativos, lo cual le condujo a posteriores investigaciones en el dominio de la teoría topológica general y clásica de los grupos continuos de Lie, campo que se ha visto completamente transformado por el trabajo de Pontriagin. Posteriormente, Pontriagin y sus discípulos llevaron a cabo una serie de estudios sobre la topología de variedades y sus aplicaciones continuas (Z. G. Boltianski, M. M. Postnikov y otros). En estos estudios se aplicó un nuevo método, la llamada ∇ -homología (cohomología), introducida en la topología combinatoria por A. N. Kolmogorov e, independientemente, por Alexander. Este método, que ahora ocupa el lugar principal en toda la topología homotópica, ha permitido desarrollar la teoría de dualidad de Pontriagin en las más diversas direcciones, llevando así a los teoremas de dualidad de A. N. Kolmogorov (y Alexander), P. S. Aleksandrov y K. A. Sitnikov, que forman parte de los resultados más impor-

tantes de la topología contemporánea. El mismo método ha encontrado también importantes aplicaciones en trabajos muy recientes de L. A. Liusternik sobre el cálculo de variaciones.

§7. Espacios topológicos y métricos

Al principio de nuestra exposición hemos hablado de la adyacencia (entre las diferentes partes de una figura dada) como uno de los conceptos topológicos fundamentales, y hemos definido las transformaciones continuas como aquellas que conservan esta relación. Sin embargo, no hemos dado una definición rigurosa de este concepto fundamental; para hacerlo con suficiente generalidad hemos de hacer uso de la teoría de conjuntos. Esta será nuestra tarea en la presente sección; resolveremos finalmente el problema introduciendo el concepto de espacio topológico.

La teoría de conjuntos permitió dar al concepto de figura geométrica una amplitud y generalidad que resultaban inaccesibles a las matemáticas «clásicas». El objeto de la investigación geométrica, en particular topológica, es ahora un conjunto arbitrario de puntos, es decir, un conjunto arbitrario cuyos elementos son puntos de un espacio euclídeo n -dimensional. Entre los puntos de un espacio n -dimensional se define una *distancia*: la distancia entre dos puntos $A = (x_1, x_2, \dots, x_n)$ y $B = (y_1, y_2, \dots, y_n)$ es, por definición, igual al número no negativo

$$\varrho(A, B) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

El concepto de distancia permite definir la adyacencia, primero entre un punto y un conjunto, y después entre dos conjuntos. Decimos que A es un punto adherente a un conjunto M , si M contiene puntos cuya distancia a A es menor que cualquier número positivo prefijado. Evidentemente, todo punto de un conjunto dado es un punto adherente de él, pero puede haber puntos que siendo adherentes no pertenezcan al conjunto. Tomemos, por ejemplo, el intervalo abierto $(0,1)$ sobre la recta numérica, esto es, el conjunto de puntos comprendidos entre el cero y el uno; los puntos 0 y 1 no pertenecen al intervalo pero son adherentes a él, ya que en el intervalo existen puntos arbitrariamente próximos a cero y puntos arbitrariamente próximos a uno. Un conjunto se llama *cerrado* si contiene todos sus puntos adherentes. Por ejemplo, el intervalo $[0, 1]$ de la recta numé-

rica, es decir, el conjunto de los puntos x que satisfacen la desigualdad $0 \leq x \leq 1$, es un conjunto cerrado. Los conjuntos cerrados en el plano y en el espacio de tres o más dimensiones pueden tener una estructura extremadamente complicada; de hecho, forman el principal objeto de estudio de la topología de conjuntos en el espacio n -dimensional.

Decimos ahora que dos conjuntos P y Q son adherentes el uno al otro si al menos uno de ellos contiene puntos adherentes del otro. De aquí se deduce que dos conjuntos cerrados sólo pueden ser adherentes cuando tienen al menos un punto en común; pero, por ejemplo, los intervalos $[0,1]$ y $(1, 2)$, que no tienen puntos comunes, son adherentes, ya que el punto 1, que pertenece a $[0,1]$, es al mismo tiempo un punto adherente de $(1, 2)$. Podemos decir entonces que un conjunto R es dividido («diseccionado») por S o que S es una «sección» de R si el conjunto $R - S$ de todos los puntos de R que no pertenecen a S se puede representar como la suma de dos conjuntos no adherentes.

Así, las ideas de Lobachevski sobre adyacencia y disección de conjuntos reciben en la topología contemporánea una expresión rigurosa y sumamente general. Hemos visto ya cómo la definición de dimensión de Uryson de un conjunto arbitrario (ver la nota del §6) está fundada en estas ideas; el enunciado de esta definición es ahora completamente riguroso. Lo mismo ocurre con la definición de aplicación o transformación continua; una aplicación f de un conjunto X en otro conjunto Y se llama continua si conserva las adherencias, es decir, si el hecho de que el punto A de X es un punto adherente de un subconjunto P de X implica que la imagen $f(A)$ de A es un punto adherente de la imagen $f(P)$ del conjunto P . Finalmente, una aplicación uno a uno del conjunto X sobre el conjunto Y se dice *topológica* si tanto ella como su inversa (de Y sobre X) son continuas. Estas definiciones proporcionan una base precisa a todo lo que llevamos dicho en las primeras secciones de la presente exposición.

Sin embargo, la topología de conjuntos no se restringe al grado de generalidad que se deriva de considerar como figuras geométricas todos los conjuntos de puntos. Es natural introducir el concepto de distancia no sólo entre puntos de un espacio euclidiano, sino entre objetos que no parecen estar en absoluto relacionados con la geometría.

Consideremos, por ejemplo, el conjunto de las funciones continuas definidas sobre el intervalo $[0, 1]$. Podemos definir la distancia $\varrho(f, g)$ entre las funciones f y g como el máximo de la expresión

$|f(x) - g(x)|$ cuando x recorre el segmento $[0, 1]$. Esta «distancia» tiene todas las propiedades básicas de la distancia entre puntos del espacio: $\varrho(f, g)$ entre las dos funciones f, g es igual a cero si, y sólo si, las funciones coinciden, es decir, si $f(x) = g(x)$ para todo punto x ; además, la distancia es evidentemente simétrica, esto es, $\varrho(f, g) = \varrho(g, f)$; finalmente, satisface la llamada ley triangular: para cada tres funciones f_1, f_2, f_3 se tiene $\varrho(f_1, f_2) + \varrho(f_2, f_3) \geq \varrho(f_1, f_3)$. Es costumbre decir que la distancia así definida convierte nuestro espacio en un espacio métrico (denotado usualmente por C). En general, entendemos por espacio métrico un conjunto de objetos arbitrarios, llamados *puntos* del espacio, entre los que se ha definido una *distancia*, número no negativo asociado a cada dos puntos que cumplen los «axiomas de distancia» antes citados.

En un espacio métrico podemos hablar de puntos adherentes a uno de sus subconjuntos y, en consecuencia, de adherencia de subconjuntos y de conceptos topológicos en general (conjuntos cerrados, aplicaciones continuas y otros conceptos que se introducen a partir de éstos). Este camino abre un extenso y muy fructífero campo de aplicación de las ideas topológicas y geométricas en general a clases de objetos matemáticos en los que hubiera parecido completamente imposible hablar de ningún tipo de geometría. Demos un ejemplo ilustrativo.

Tomemos de nuevo la ecuación diferencial (2)

$$\frac{dy}{dx} = F(x, y).$$

Si $y = \phi(x)$ para $0 \leq x \leq 1$ es una solución de la ecuación que toma para $x = 0$ el valor $y = 0$, entonces la función $\phi(x)$ satisface evidentemente la ecuación integral

$$\phi(x) = \int_0^x F[x, \phi(x)] dx. \quad (3)$$

Consideremos ahora la integral $G(f) = \int_0^x F[x, f(x)] dx$, donde $0 \leq x \leq 1$ y $f(x)$ es una función continua arbitraria definida en el intervalo $[0, 1]$. La expresión $G(f) = \int_0^x F[x, f(x)] dx$ asocia así a cada función f una función $g = G(f)$, ya que la integral es una función continua $g(x)$ también definida en $[0, 1]$; en otros términos,

tenemos una aplicación continua G del espacio métrico C en sí mismo. ¿Cómo podemos caracterizar una función $\phi(x)$ (puede haber varias) que sea solución de la ecuación (2) o de su equivalente (3)? Evidentemente, nuestra aplicación la transforma en ella misma; es decir, es un punto fijo de la aplicación G . Resulta ahora que tal punto fijo de la aplicación G efectivamente existe, como se deduce de un teorema muy general sobre puntos fijos de aplicaciones continuas en espacios métricos probado en 1926 por P. S. Aleksandrov y V. V. Nemyski. En nuestros días, el estudio de diversos espacios métricos cuyos puntos son funciones de una u otra clase (tales espacios se llaman funcionales) es un instrumento de uso constante en análisis; el estudio de los espacios funcionales por métodos que son en parte topológicos, pero principalmente algebraicos, en el sentido amplio de la palabra, forma el contenido del análisis funcional (ver capítulo 18).

El análisis funcional, como ya mencionamos en el capítulo introductorio, ocupa un lugar sobresaliente en la matemática contemporánea en razón de la variedad de sus relaciones con numerosas partes de las matemáticas y de su valor en las ciencias naturales, sobre todo en física teórica. La investigación de las propiedades topológicas de los espacios funcionales está íntimamente relacionada con el cálculo de variaciones y la teoría de ecuaciones en derivadas parciales (trabajos de Liusternik, Morse, Leray, Schauder, Krasnoselski y otros). Los problemas sobre la existencia de puntos fijos para aplicaciones continuas de espacios funcionales juegan un importante papel en esta investigación.

La topología de espacios métricos generales y espacios funcionales no es la última palabra en lo que respecta a la generalidad de las teorías topológicas contemporáneas. El hecho es que en un espacio métrico el concepto fundamental de adherencia se introduce a partir de la distancia entre dos puntos, que por el contrario no es un concepto topológico. El problema reside por tanto en una definición directa y axiomática de adherencia. De esta forma se llega al concepto de espacio topológico, el concepto más general de la topología actual.

Un *espacio topológico* es un conjunto de objetos de naturaleza arbitraria (que se denominan puntos del espacio) en el cual los puntos adherentes de cada subconjunto están dados de alguna forma. Se supone además que se satisfacen ciertas condiciones naturales, llamadas axiomas del espacio topológico (por ejemplo, todo punto de un conjunto es un punto adherente del mismo, un punto adherente

de la suma de dos conjuntos es un punto adherente de al menos uno de los conjuntos, etc.). Hoy día se están llevando a cabo profundos trabajos en la teoría de espacios topológicos; en su desarrollo los matemáticos soviéticos P. S. Uryzon, P. S. Aleksandrov, A. N. Tihonov y otros desempeñan un papel preponderante. De los últimos resultados debemos mencionar uno de valor fundamental: el joven matemático Ju. M. Smirnov ha encontrado condiciones necesarias y suficientes para la metrizableidad de un espacio topológico, es decir, condiciones bajo las cuales se puede definir una distancia entre los puntos del espacio, de modo que la topología del espacio se puede considerar como engendrada por esa distancia; dicho con otras palabras, tal que los puntos adherentes de todos los posibles conjuntos del espacio métrico obtenido son los mismos que los definidos inicialmente en el espacio topológico dado.

El surgimiento y difusión del análisis funcional en el siglo xx tiene dos causas principales. Por una parte era deseable interpretar desde un punto de vista uniforme el copioso material acumulado a lo largo del siglo xix en distintas, y a menudo casi inconexas, ramas de las matemáticas. Los conceptos fundamentales del análisis funcional se formaron y cristalizaron bajo varios aspectos y por distintas razones. Muchos de sus conceptos fundamentales surgieron de forma natural en el proceso de desarrollo del cálculo de variaciones, en problemas sobre oscilaciones (en la transición de las oscilaciones de un sistema con un número finito de grados de libertad a oscilaciones de un medio continuo), en la teoría de ecuaciones integrales, en la teoría de ecuaciones diferenciales ordinarias y en derivadas parciales (en problemas de contorno, problemas de autovalores, etcétera), en la teoría de funciones de una variable real, en el cálculo de operadores, en la discusión de problemas de la teoría de aproximación de funciones y otros. El análisis funcional posibilitó la comprensión de muchos resultados de esos campos desde un único punto de vista y promovió la creación de otros nuevos. En las últimas décadas, los conceptos y el aparato matemático del análisis funcional han sido utilizados en una nueva rama de la física teórica: la mecánica cuántica.

Por otra parte, el estudio de los problemas matemáticos rela-

cionados con la mecánica cuántica se convirtió en un hecho crucial en el desarrollo posterior del análisis funcional: creó, y todavía crea en la actualidad, ramas fundamentales de su desarrollo.

El análisis funcional no ha llegado, ni mucho menos, a su total realización. Por el contrario, es indudable que en su futuro desarrollo los problemas y necesidades de la física contemporánea tendrán para él tanta importancia como la mecánica clásica para el nacimiento y desarrollo del cálculo diferencial e integral en el siglo XVIII.

Es imposible incluir en este capítulo todos los problemas del análisis funcional, ni siquiera los más fundamentales. Muchas de sus importantes ramas sobrepasan los límites de este libro. No obstante, limitándonos a ciertos problemas selectos, deseamos poner al lector al corriente de los conceptos fundamentales del análisis funcional, ilustrando, en la medida de lo posible, las relaciones de las que antes hablábamos. Estos problemas fueron analizados a comienzos del siglo XX sobre la base de los trabajos clásicos de Hilbert, que fue uno de los fundadores del análisis funcional. Desde entonces, esta disciplina se ha desarrollado vigorosamente y ha conocido aplicaciones a casi todas las ramas de las matemáticas; en ecuaciones diferenciales, en derivadas parciales, en la teoría de probabilidades, en mecánica cuántica, en la teoría cuántica de campos, etc. Desgraciadamente, estos últimos desarrollos del análisis funcional no pueden ser incluidos en nuestra exposición. Describirlos exigiría por sí solo un extenso volumen. Nos restringiremos, por tanto, a uno de los problemas más antiguos: la teoría de autofunciones.

§1. *El espacio n -dimensional*

En lo que sigue haremos uso de los conceptos fundamentales del espacio n -dimensional. Aunque estos conceptos han sido introducidos en los capítulos sobre álgebra lineal y espacios abstractos, no creemos que sea superfluo repetirlos en la forma en que aparecen en el análisis funcional. Para entender esta sección es suficiente que el lector posea una idea de los fundamentos de la geometría analítica.

Sabemos que en la geometría analítica del espacio tridimensional un punto está determinado por una terna de números (f_1, f_2, f_3) , que son sus coordenadas. La distancia de este punto al origen de coordenadas es igual a $\sqrt{f_1^2 + f_2^2 + f_3^2}$. Si consideramos un punto como el extremo del vector que lo une al origen de coordenadas, entonces la

longitud de este vector es también igual a $\sqrt{f_1^2 + f_2^2 + f_3^2}$. El ángulo entre dos vectores que unen el origen de coordenadas con los puntos $A(f_1, f_2, f_3)$ y $B(g_1, g_2, g_3)$ tiene por coseno el número definido por la fórmula

$$\cos \phi = \frac{f_1 g_1 + f_2 g_2 + f_3 g_3}{\sqrt{f_1^2 + f_2^2 + f_3^2} \sqrt{g_1^2 + g_2^2 + g_3^2}}.$$

Sabemos por trigonometría que $|\cos \phi| \leq 1$. Tenemos, por tanto, la desigualdad

$$\frac{f_1 g_1 + f_2 g_2 + f_3 g_3}{\sqrt{f_1^2 + f_2^2 + f_3^2} \sqrt{g_1^2 + g_2^2 + g_3^2}} \leq 1,$$

y por tanto, también

$$(f_1 g_1 + f_2 g_2 + f_3 g_3)^2 \leq (f_1^2 + f_2^2 + f_3^2)(g_1^2 + g_2^2 + g_3^2). \quad (1)$$

Esta última desigualdad es de carácter algebraico y es válida para cualesquiera seis números (f_1, f_2, f_3) y (g_1, g_2, g_3) , puesto que todo conjunto de seis números se puede considerar siempre como las coordenadas de dos puntos del espacio. Sea como fuere, la desigualdad (1) se obtuvo a través de consideraciones puramente geométricas y está estrechamente relacionada con la geometría, lo cual permite visualizar fácilmente su significado.

En la formulación analítica de una serie de relaciones geométricas sucede con frecuencia que los hechos correspondientes siguen siendo verdaderos cuando se reemplaza la terna de números por n números. Por ejemplo, nuestra desigualdad (1) se puede generalizar a $2n$ números $(f_1, f_2, \dots, f_n)$ y $(g_1, g_2, \dots, g_n)$. Esto significa que para cada $2n$ números $(f_1, f_2, \dots, f_n)$ y $(g_1, g_2, \dots, g_n)$ la siguiente desigualdad, análoga a la (1), sigue siendo verdadera:

$$(f_1 g_1 + f_2 g_2 + \dots + f_n g_n)^2 \leq (f_1^2 + f_2^2 + \dots + f_n^2)(g_1^2 + g_2^2 + \dots + g_n^2) \quad (1')$$

Esta desigualdad, de la que (1) es un caso especial, se puede probar por métodos puramente analíticos¹. De forma similar, muchas otras

¹ Ver capítulo 15.

relaciones entre ternas de números derivadas de la geometría analítica se pueden generalizar a n números. Esta conexión de la geometría con las relaciones entre números (relaciones numéricas), de la que es ejemplo la citada desigualdad, se hace particularmente clara cuando se introduce el concepto del espacio n -dimensional. Dicho concepto fue introducido en el capítulo 15. Lo repetimos aquí brevemente.

Una colección de n números $(f_1, f_2, \dots, f_n)$ es un punto o *vector* de un espacio n -dimensional (usaremos más a menudo el último nombre). Designaremos el vector $(f_1, f_2, \dots, f_n)$ por la única letra f .

Del mismo modo que en el espacio tridimensional la suma de vectores se realiza sumando sus componentes, definimos la suma de los vectores

$$f = \{f_1, f_2, \dots, f_n\} \text{ y } g = \{g_1, g_2, \dots, g_n\}$$

como el vector $\{f_1 + g_1, f_2 + g_2, \dots, f_n + g_n\}$ y lo denotamos por $f + g$.

El producto del vector $f = \{f_1, f_2, \dots, f_n\}$ por el número λ es el vector $\lambda f = \{\lambda f_1, \lambda f_2, \dots, \lambda f_n\}$.

La longitud del vector $f = \{f_1, f_2, \dots, f_n\}$, al igual que la de un vector en un espacio tridimensional, se define por el número $\sqrt{f_1^2 + f_2^2 + \dots + f_n^2}$.

El ángulo ϕ entre los dos vectores $f = \{f_1, f_2, \dots, f_n\}$ y $g = \{g_1, g_2, \dots, g_n\}$ en un espacio n -dimensional viene dado por su coseno de la misma forma que en el espacio tridimensional. Se define por la fórmula ²

$$\cos \phi = \frac{f_1 g_1 + f_2 g_2 + \dots + f_n g_n}{\sqrt{f_1^2 + f_2^2 + \dots + f_n^2} \sqrt{g_1^2 + g_2^2 + \dots + g_n^2}}. \quad (2)$$

El producto escalar de dos vectores es el nombre que se da al producto de sus longitudes por el coseno del ángulo que forman. Así, si $f = \{f_1, f_2, \dots, f_n\}$ y $g = \{g_1, g_2, \dots, g_n\}$, como sus longitudes son respectivamente $\sqrt{f_1^2 + f_2^2 + \dots + f_n^2}$ y $\sqrt{g_1^2 + g_2^2 + \dots + g_n^2}$, su producto escalar, que denotamos por (f, g) , viene dado por la fórmula

$$(f, g) = f_1 g_1 + f_2 g_2 + \dots + f_n g_n. \quad (3)$$

En particular, la condición de ortogonalidad (perpendicularidad) de dos vectores es la ecuación $\cos \phi = 0$; es decir, $(f, g) = 0$.

² La desigualdad $|\cos \phi| \leq 1$, se sigue de la desigualdad (1').

Por medio de la fórmula (3) el lector puede verificar que el producto escalar en un espacio n -dimensional tiene las siguientes propiedades:

1. $(f, g) = (g, f)$.
2. $(\lambda f, g) = \lambda(f, g)$.
3. $(f, g_1 + g_2) = (f, g_1) + (f, g_2)$.
4. $(f, f) \geq 0$, y la igualdad se verifica sólo para $f = 0$, es decir, cuando $f_1 = f_2 = \dots = f_n = 0$.

El producto escalar de un vector por sí mismo (f, f) es igual al cuadrado de la longitud de f .

El producto escalar es un instrumento muy conveniente en el estudio de los espacios n -dimensionales. No estudiaremos aquí la geometría de un espacio n -dimensional; nos limitaremos a un solo ejemplo.

Tomemos el teorema de Pitágoras en un espacio n -dimensional: el cuadrado de la hipotenusa es igual a la suma de los cuadrados de los catetos. La demostración que damos de este teorema en el plano se puede trasladar fácilmente al caso del espacio n -dimensional.

Sean f y g dos vectores perpendiculares en el plano. Consideremos el triángulo rectángulo construido sobre f y g (Fig. 1). La hipotenusa de este triángulo es igual a la longitud del vector $f + g$. Escribamos el teorema de Pitágoras en forma vectorial con nuestra notación. Puesto que el cuadrado de la longitud de un vector es igual al producto escalar del vector por sí mismo, el teorema de Pitágoras se puede escribir en el lenguaje de productos escalares como sigue:

$$(f + g, f + g) = (f, f) + (g, g).$$

La demostración se sigue inmediatamente de las propiedades del producto escalar. Sabemos que

$$(f + g, f + g) = (f, f) + (f, g) + (g, f) + (g, g),$$

y los dos sumandos (f, g) y (g, f) son cero debido a la ortogonalidad de f y g .

En esta demostración hemos usado solamente la definición de longitud de un vector, de perpendicularidad de vectores y las propiedades del producto escalar. Por tanto, nada cambia en la demostración cuando suponemos que f y g son dos vectores ortogonales de un

espacio n -dimensional. Y así se prueba el teorema de Pitágoras para un triángulo rectángulo en un espacio n -dimensional.

Dados tres vectores ortogonales dos a dos f, g y h en un espacio n -dimensional, su suma $f + g + h$ es la diagonal del paralelepípedo

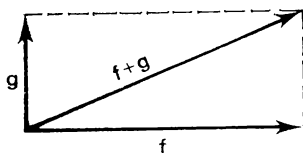


FIG. 1

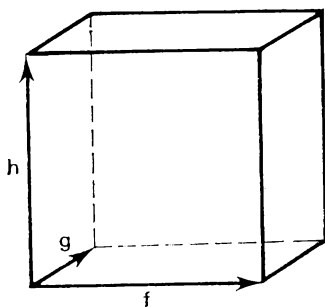


FIG. 2

recto construido con esos vectores (Fig. 2); tenemos así la ecuación siguiente:

$$(f + g + h, f + g + h) = (f, f) + (g, g) + (h, h),$$

lo que significa que el cuadrado de la diagonal de un paralelepípedo es igual a la suma de los cuadrados de las longitudes de sus aristas. La demostración de esta aserción, que es del todo análoga a la dada anteriormente para el teorema de Pitágoras, se deja al lector. Análogamente, si en un espacio n -dimensional hay k vectores ortogonales dos a dos, $f^1, f^2, \dots, f^k$, entonces la ecuación

$$(f^1 + f^2 + \dots + f^k, f^1 + f^2 + \dots + f^k) = (f^1, f^1) + (f^2, f^2) + \dots + (f^k, f^k), \quad (4)$$

que se prueba de la misma forma, significa que el cuadrado de la longitud de la diagonal de un «paralelepípedo k -dimensional» en un espacio n -dimensional es también igual a la suma de los cuadrados de las longitudes de sus aristas.

§2. *Espacio de Hilbert (Espacio infinito-dimensional)**Conexión con el espacio n -dimensional*

La introducción del concepto de espacio n -dimensional resultó ser de gran provecho en el estudio de muchos problemas de matemáticas y física. A su vez, este concepto impulsó el ulterior desarrollo del concepto de espacio y sus aplicaciones en los distintos dominios de las matemáticas. Los problemas de pequeñas oscilaciones de sistemas elásticos jugaron un importante papel en el desarrollo del álgebra lineal y de la geometría del espacio n -dimensional.

Consideremos el siguiente ejemplo clásico (Fig. 3). Sea AB una cuerda flexible tensada entre los puntos A y B . Supongamos que sujetamos un cierto peso en un punto C de la misma. Si lo desplazamos de su posición inicial de equilibrio, comienza a oscilar con una cierta frecuencia ω , que podemos calcular si conocemos la tensión de la cuerda, la masa m y la posición del peso. El estado del sistema en cada instante está entonces determinado por un único número: el desplazamiento y_1 de la masa m a partir de la posición de equilibrio de la cuerda.

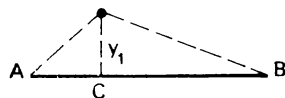


FIG. 3

Coloquemos ahora n pesos sobre la cuerda AB en los puntos $C_1, C_2, \dots, C_n$. Suponemos que la cuerda AB en sí no tiene peso. Esto significa que su masa es tan pequeña que, comparada con las masas de los pesos, se puede despreciar. El estado del sistema viene dado por n números $y_1, y_2, \dots, y_n$, iguales a los desplazamientos de los pesos a partir de su posición de equilibrio. La colección de números $y_1, y_2, \dots, y_n$ se puede considerar (y esto resulta muy útil en muchos aspectos) como un vector $(y_1, y_2, \dots, y_n)$ de un espacio n -dimensional.

El estudio de las pequeñas oscilaciones que tienen lugar en estas circunstancias está estrechamente conectado con las propiedades de la geometría en los espacios n -dimensionales. Se puede probar, por ejemplo, que la determinación de la frecuencia de las oscilaciones de un sistema tal se reduce a la tarea de hallar los ejes de un cierto elipsoide en el espacio n -dimensional.

Consideremos ahora el problema de las pequeñas oscilaciones de una cuerda tensada entre los puntos A y B . Aquí estamos pensando en una cuerda ideal, esto es, un hilo elástico de masa finita

distribuida continuamente a lo largo de su longitud. En particular, entendemos por cuerda homogénea una cuerda de densidad constante.

Puesto que la masa está distribuida continuamente a lo largo de la cuerda, su posición no se puede ya dar por un conjunto finito de números, sino que se debe especificar el desplazamiento $y(x)$ de cada punto x de la cuerda. De esta forma el estado de la cuerda en cada instante se describe por una cierta función $y(x)$.

El estado de un hilo con n pesos fijados a los puntos de coordenadas $x_1, x_2, \dots, x_n$ se

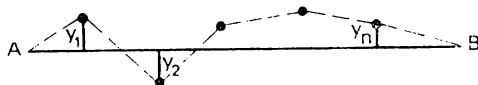


FIG. 4

representa gráficamente por una línea quebrada de n componentes (figura 4), de suerte que, cuando aumenta el número de pesos, el número de segmentos de la

línea quebrada aumenta también de modo correspondiente. Cuando el número de pesos crece indefinidamente y la distancia entre dos adyacentes tiende a cero, obtenemos, en el límite, una distribución continua de masas a lo largo del hilo, es decir, una cuerda idealizada. La línea quebrada que describe la posición del hilo con pesos se convierte entonces en una curva que determina la posición de la cuerda (Fig. 5).

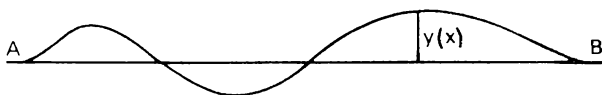


FIG. 5

Vemos así que existe una estrecha conexión entre las oscilaciones de un hilo con pesos y las oscilaciones de una cuerda. En el primer problema, la posición del sistema estaba dada por un punto o vector en un espacio n -dimensional. Por tanto es natural considerar la función $f(x)$ que describe la posición de una cuerda oscilante en el segundo caso como un punto o un vector de un espacio infinito-dimensional. Toda una serie de problemas similares conducen a la misma idea de considerar espacios cuyos puntos (vectores) son funciones definidas sobre cierto intervalo³.

³ Podemos considerar también como uno de tales problemas el de las oscilaciones eléctricas que se establecen en una serie de circuitos conectados (Fig. 6).

Este ejemplo de las oscilaciones de una cuerda sobre el que volveremos en la sección 4, sugiere la manera de introducir los conceptos fundamentales en un espacio infinito-dimensional.

El espacio de Hilbert

Discutiremos ahora el concepto de espacio de Hilbert, uno de los ejemplos de espacio infinito-dimensional más extendidos y de la máxima importancia en las aplicaciones.

Un vector de un espacio n -dimensional se define como una colección de n números f_i , donde i varía de 1 a n . De modo similar, un vector de un espacio infinito-dimensional se define como una función $f(x)$, donde x varía entre a y b .

La adición de vectores y la multiplicación por un número se define igual que la adición de funciones y la multiplicación de una función por un número.

La longitud de un vector f en un espacio n -dimensional se define por la fórmula

$$\sqrt{\sum_{i=1}^n f_i^2}.$$

Puesto que para funciones las integrales juegan el papel de la suma, la longitud de un vector $f(x)$ en un espacio de Hilbert se da mediante la fórmula

$$\sqrt{\int_a^b f^2(x) dx}. \quad (5)$$

El estado de esta serie se puede expresar por un conjunto de n números $u_1, u_2, \dots, u_n$ donde u_i es el voltaje del condensador del circuito i de la cadena. La colección de n números $(u_1, u_2, \dots, u_n)$ es un vector en el espacio n -dimensional.

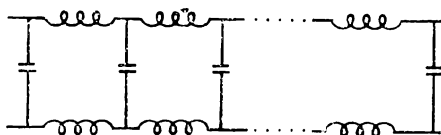


FIG. 6

Imaginemos ahora una línea formada por dos conductores de capacidad e inductancia finitas, distribuidas a lo largo de la línea. Su estado eléctrico se expresa mediante una cierta función $u(x)$, que da la distribución del voltaje a lo largo de la línea. Esta función es un vector del espacio de dimensión infinita de las funciones definidas en el intervalo (a, b) .

La distancia entre los puntos f y g de un espacio n -dimensional se define como la longitud del vector $f - g$, esto es, por la fórmula

$$\sqrt{\sum_{i=1}^n (f_i - g_i)^2},$$

Análogamente, la «distancia» entre los elementos $f(t)$ y $g(t)$ en un espacio funcional es igual a

$$\sqrt{\int_a^b [f(t) - g(t)]^2 dt}$$

La expresión $\int_a^b [f(t) - g(t)]^2 dt$ se llama desviación en media cuadrática de las funciones $f(t)$ y $g(t)$. Así, pues, la desviación en media cuadrática de dos elementos de un espacio de Hilbert es la medida de sus distancias.

Procedamos ahora a la definición de ángulo entre vectores. En un espacio n -dimensional, el ángulo ϕ entre los vectores $f = \{f_i\}$ y $g = \{g_i\}$ se define por la fórmula

$$\cos \phi = \frac{\sum_{i=1}^n f_i g_i}{\sqrt{\sum_{i=1}^n f_i^2} \sqrt{\sum_{i=1}^n g_i^2}}.$$

En un espacio infinito-dimensional, las sumas se sustituyen por las integrales correspondientes, y el ángulo ϕ entre los dos vectores f y g de un espacio de Hilbert se define por la fórmula siguiente:

$$\cos \phi = \frac{\int_a^b f(t) g(t) dt}{\sqrt{\int_a^b f^2(t) dt} \sqrt{\int_a^b g^2(t) dt}}. \quad (6)$$

Esta expresión se puede considerar como el coseno de un ángulo siempre que la fracción de la derecha sea en valor absoluto menor que 1, es decir, si

$$\left| \int_a^b f(t) g(t) dt \right| < \sqrt{\int_a^b f^2(t) dt} \sqrt{\int_a^b g^2(t) dt}. \quad (7)$$

Esta desigualdad se cumple de hecho para dos funciones arbitrarias $f(t)$ y $g(t)$. Juega un importante papel en análisis y se conoce con el nombre de desigualdad de Cauchy-Bunjakovski. Probémosla.

Sean $f(x)$ y $g(x)$ dos funciones, no idénticamente nulas, definidas en un intervalo (a, b) . Tomamos dos números arbitrarios λ y μ y formamos la expresión

$$\int_a^b [\lambda f(x) - \mu g(x)]^2 dx.$$

Puesto que la función $[\lambda f(x) - \mu g(x)]^2$ bajo el signo integral es no negativa, tenemos la siguiente desigualdad:

$$\int_a^b [\lambda f(x) - \mu g(x)]^2 dx \geq 0;$$

es decir,

$$2\lambda\mu \int_a^b f(x) g(x) dx \leq \lambda^2 \int_a^b f^2(x) dx + \mu^2 \int_a^b g^2(x) dx.$$

Por brevedad introducimos la notación

$$\left| \int_a^b f(x) g(x) dx \right| = C, \quad \int_a^b f^2(x) dx = A, \quad \int_a^b g^2(x) dx = B. \quad (8)$$

Con esta notación, la desigualdad se puede volver a escribir como sigue ⁴:

$$2\lambda\mu C \leq \lambda^2 A + \mu^2 B. \quad (9)$$

Esta desigualdad es válida para valores arbitrarios de λ y μ ; en particular podemos tomar

$$\lambda = \sqrt{\frac{C}{A}}, \quad \mu = \sqrt{\frac{C}{B}}. \quad (10)$$

Substituyendo estos valores de λ y μ en (9), obtenemos

$$\frac{C}{\sqrt{AB}} \leq 1.$$

Si sustituimos A , B y C por sus expresiones en (8), obtenemos finalmente la desigualdad de Cauchy-Bunjakovski.

⁴ Debemos tomar para C el módulo de la integral debido al signo arbitrario de λ y μ .

En geometría el producto escalar de dos vectores se define como el producto de sus longitudes por el coseno del ángulo que forman. Las longitudes de los vectores f y g son en nuestro caso iguales a

$$\sqrt{\int_a^b f^2(x) dx} \quad \text{y} \quad \sqrt{\int_a^b g^2(x) dx},$$

y el coseno del ángulo que forman se define por la fórmula

$$\cos \phi = \frac{\int_a^b f(x) g(x) dx}{\sqrt{\int_a^b f^2(x) dx} \sqrt{\int_a^b g^2(x) dx}}.$$

Si multiplicamos estas expresiones, llegamos a la siguiente fórmula para el producto escalar de dos vectores en un espacio de Hilbert:

$$(f, g) = \int_a^b f(x) g(x) dx. \quad (11)$$

A partir de esta fórmula se ve claramente que el producto escalar del vector f por sí mismo es el cuadrado de su longitud.

Si el producto escalar de dos vectores distintos de cero es igual a cero, ello se debe a que $\cos \phi = 0$, es decir, a que el ángulo que les corresponde según nuestra definición es 90° . Por tanto, las funciones f y g para las que

$$(f, g) = \int_a^b f(x) g(x) dx = 0$$

se llaman ortogonales.

El teorema de Pitágoras (ver §1) se verifica en un espacio de Hilbert como en un espacio n -dimensional. Sean $f_1(x), f_2(x), \dots, f_N(x)$ N funciones ortogonales dos a dos y

$$f(x) = f_1(x) + f_2(x) + \dots + f_N(x).$$

Entonces el cuadrado de la longitud de f es igual a la suma de los cuadrados de las longitudes de $f_1, f_2, \dots, f_N$.

Puesto que las longitudes de vectores en un espacio de Hilbert se

dan por medio de integrales, el teorema de Pitágoras en este caso se expresa por la fórmula

$$\int_a^b f^2(x) dx = \int_a^b f_1^2(x) dx + \int_a^b f_2^2(x) dx + \cdots + \int_a^b f_N^2(x) dx \quad (12)$$

La demostración de este teorema no difiere en ningún aspecto de la dada anteriormente (§1) para el mismo teorema en el espacio n -dimensional.

Hasta ahora no hemos precisado qué funciones se deben considerar como vectores de un espacio de Hilbert. Como tales funciones hay que tomar aquellas para las que $\int_a^b f^2(x) dx$ tiene sentido. Parece natural limitarnos a funciones continuas para las que $\int_a^b f^2(x) dx$ exista siempre. Sin embargo, la teoría de Hilbert resulta más completa y natural si $\int_a^b f^2(x)$ se interpreta en un sentido generalizado, esto es, como una integral de Lebesgue (ver capítulo 15).

Esta extensión del concepto de integral (y por consiguiente de la clase de funciones a considerar) es tan necesaria para el análisis funcional como lo es para los fundamentos del cálculo diferencial e integral una teoría rigurosa de los números reales. Así, la generalización del concepto ordinario de integral, creado a comienzos del siglo xx en conexión con el desarrollo de la teoría de funciones de una variable real, resultó ser esencial para el análisis funcional y las ramas de la matemática conectadas con él.

§3. Desarrollo mediante sistemas ortogonales de funciones

Definición y ejemplos de sistemas ortogonales de funciones

Si tomamos en un plano dos vectores arbitrarios e_1, e_2 perpendiculares entre sí y de longitud unidad (figura 7), todo vector del plano se puede descomponer en la dirección de esos dos vectores, es decir, se puede representar en la forma

$$f = a_1 e_1 + a_2 e_2,$$

donde a_1, a_2 son números iguales a las proyecciones del vector f sobre las direcciones de los ejes de e_1 y e_2 . Puesto que la proyección de f sobre un eje es igual al producto de la longitud de f

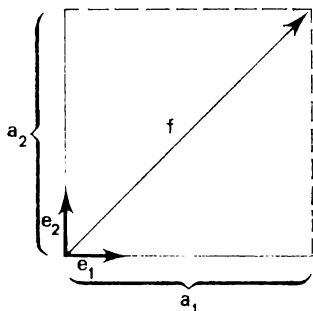


FIG. 7

por el coseno del ángulo que forma con el eje, podemos escribir, recordando la definición de producto escalar,

$$\begin{aligned} a_1 &= (f, e_1), \\ a_2 &= (f, e_2). \end{aligned}$$

Análogamente, si en un espacio tridimensional tomamos tres vectores e_1, e_2, e_3 , unitarios y perpendiculares dos a dos, entonces todo vector f de este espacio se puede escribir en la forma

$$f = a_1 e_1 + a_2 e_2 + a_3 e_3,$$

donde

$$a_k = (f, e_k) \quad (k = 1, 2, 3).$$

En un espacio de Hilbert podemos también considerar sistemas de vectores ortogonales dos a dos, es decir, funciones $\phi_1(x), \phi_2(x), \dots, \phi_n(x), \dots$.

Tales sistemas de funciones juegan un importante papel en análisis y se llaman sistemas ortogonales. Aparecen en muy diversos problemas de matemática física, ecuaciones integrales, cálculo aproximado en la teoría de funciones de una variable real, etc. La ordenación y unificación de los conceptos relativos a tales sistemas fue uno de los motivos que condujeron, a comienzos del siglo XX, a la creación del concepto general de espacio de Hilbert.

Demos una definición precisa. Un sistema de funciones

$$\phi_1(x), \phi_2(x), \dots, \phi_n(x), \dots$$

se llama *ortogonal* si cada dos funciones del sistema son ortogonales, es decir, si

$$\int_a^b \phi_i(x) \phi_k(x) dx = 0 \quad \text{para } i \neq k. \quad (13)$$

En el espacio tridimensional exigíamos que los vectores del sistema tuvieran longitud unidad. Recordando la definición de longitud en un espacio de Hilbert podemos escribir, en nuestro caso, la condición como sigue:

$$\int_a^b \phi_k^2(x) dx = 1. \quad (14)$$

Un sistema que satisface las condiciones (13) y (14) se llama **ortonormal**.

Demos ejemplos de tales sistemas de funciones.

1. Consideramos sobre el intervalo $(-\pi, \pi)$ la sucesión de funciones

$$1, \cos x, \sin x, \cos 2x, \sin 2x, \dots, \cos nx, \sin nx, \dots$$

Dos funciones cualesquiera de esta sucesión son ortogonales entre sí. Esto se puede verificar mediante un simple cálculo de las correspondientes integrales. El cuadrado de la longitud de un vector en un espacio de Hilbert es la integral del cuadrado de la función. Así, pues, las longitudes de los vectores de la sucesión, al cuadrado, son las integrales

$$\int_{-\pi}^{\pi} dx = 2\pi, \int_{-\pi}^{\pi} \cos^2 nx \, dx = \pi, \int_{-\pi}^{\pi} \sin^2 nx \, dx = \pi,$$

es decir, los vectores de nuestra sucesión son ortogonales, pero no están normalizados. La longitud del primer vector de la sucesión es $\sqrt{2\pi}$, y la de todos los demás $\sqrt{\pi}$. Si dividimos cada vector por su longitud obtenemos la sucesión de funciones trigonométricas

$$\frac{1}{\sqrt{2\pi}}, \frac{\cos x}{\sqrt{\pi}}, \frac{\sin x}{\sqrt{\pi}}, \frac{\cos 2x}{\sqrt{\pi}}, \frac{\sin 2x}{\sqrt{\pi}}, \dots, \frac{\cos nx}{\sqrt{\pi}}, \frac{\sin nx}{\sqrt{\pi}}, \dots$$

Este sistema es históricamente uno de los primeros y más importantes ejemplos de sistemas ortogonales. Aparece en los trabajos de Euler, D. Bernouilli, y d'Alembert en relación con problemas de oscilaciones de cuerdas. Su estudio juega un papel esencial en el desarrollo de todo el análisis ⁵.

La aparición del sistema ortogonal de funciones trigonométricas en relación con los problemas de oscilaciones de cuerdas no es accidental. Todo problema de pequeñas oscilaciones de un medio conduce a un cierto sistema de funciones ortogonales que describe las llamadas oscilaciones características del sistema dado (ver §4). Por ejemplo, en relación con las oscilaciones de una esfera aparecen las llamadas funciones esféricas; en relación con los problemas de osci-

⁵ Ver capítulo 12, §1.

laciones de una membrana circular o un cilindro aparecen las llamadas funciones cilíndricas, etc.

2. Podemos dar un ejemplo de un sistema ortogonal de funciones formado por polinomios. Tal ejemplo es la sucesión de polinomios de Legendre

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n (x^2 - 1)^n}{dx^n}.$$

es decir, $P_n(x)$ (salvo un factor constante) es la derivada n -ésima de $(x^2 - 1)^n$. Escribamos los cuatro primeros polinomios de la sucesión:

$$\begin{aligned} P_0(x) &= 1; \\ P_1(x) &= x; \\ P_2(x) &= \frac{1}{2}(3x^2 - 1); \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x). \end{aligned}$$

Es obvio que $P_n(x)$ es un polinomio de grado n . Dejamos al lector que demuestre que estos polinomios forman una sucesión ortogonal sobre el intervalo $(-1, 1)$.

La teoría general de polinomios ortogonales fue desarrollada en la segunda mitad del siglo XIX por el famoso matemático ruso P. L. Chebichev.

Desarrollos mediante sistemas ortogonales de funciones

De la misma forma que en un espacio tridimensional todo vector se puede representar en forma de una combinación lineal de tres vectores ortogonales dos a dos y de longitud unidad

$$f = a_1 e_1 + a_2 e_2 + a_3 e_3,$$

así, en un espacio funcional, surge el problema de la descomposición de una función arbitraria f en una serie respecto a un sistema ortogonal de funciones, es decir, el problema de la representación de f en la forma

$$f(x) = a_1 \phi_1(x) + a_2 \phi_2(x) + \cdots + a_n \phi_n(x) + \cdots \quad (15)$$

Aquí, la convergencia de la serie (15) hacia la función f se debe entender en el sentido de la distancia entre elementos de un espacio

de Hilbert. Esto significa que la desviación en media cuadrática de la suma parcial de la serie

$$S_n(t) = \sum_{k=1}^n a_k \phi_k(t)$$

respecto a la función $f(t)$ tiende a cero cuando $n \rightarrow \infty$; es decir,

$$\lim_{n \rightarrow \infty} \int_a^b [f(t) - S_n(t)]^2 dt = 0. \quad (16)$$

Esta convergencia se suele llamar «convergencia en media».

Los desarrollos de funciones en distintos sistemas de funciones ortogonales aparecen a menudo en análisis y son un importante método para la solución de problemas de la física matemática. Por ejemplo, si el sistema ortogonal es el de las funciones trigonométricas sobre el intervalo $(-\pi, \pi)$

$$1, \cos x, \operatorname{sen} x, \cos 2x, \operatorname{sen} 2x, \dots, \cos nx, \operatorname{sen} nx, \dots,$$

entonces el desarrollo es el clásico desarrollo de una función en serie trigonométrica ⁶

$$f(x) = a_0 + a_1 \cos x + b_1 \operatorname{sen} x + a_2 \cos 2x + b_2 \operatorname{sen} 2x + \dots$$

Supongamos que es posible un desarrollo como el (15) para toda función f de un espacio de Hilbert y tratemos de encontrar sus coeficientes a_n . Para ello multiplicamos ambos miembros de la ecuación escalarmente por una de las funciones ϕ_m de nuestro sistema. Obtenemos la ecuación

$$(f, \phi_m) = a_1(\phi_1, \phi_m) + a_2(\phi_2, \phi_m) + \dots + a_m(\phi_m, \phi_m) + a_{m+1}(\phi_{m+1}, \phi_m) + \dots,$$

y debido a que $(\phi_m, \phi_n) = 0$ para $m \neq n$ y $(\phi_m, \phi_m) = 1$, queda determinado el valor del coeficiente a_m

$$a_m = (f, \phi_m) \quad (m = 1, 2, \dots).$$

⁶ Tales descomposiciones se dan frecuentemente en distintos problemas de física en la descomposición de una oscilación en los armónicos que la constituyen. Ver capítulo 6, §5.

Vemos que, como en el espacio tridimensional ordinario (ver el comienzo de esta sección), los coeficientes a_m son iguales a las proyecciones del vector f en la dirección de los vectores ϕ_k .

Recordando la definición de producto escalar, se observa que los coeficientes del desarrollo de $f(x)$ mediante el sistema ortogonal de funciones $\phi_1(x), \phi_2(x), \dots, \phi_n(x), \dots$

$$f(x) = a_1\phi_1(x) + a_2\phi_2(x) + \dots + a_n\phi_n(x) + \dots \quad (17)$$

están determinados por las fórmulas

$$a_m = \int_a^b f(t) \phi_m(t) dt. \quad (18)$$

Consideremos, por ejemplo, el sistema trigonométrico ortonormal mencionado anteriormente:

$$\frac{1}{\sqrt{2\pi}}, \frac{\cos x}{\sqrt{\pi}}, \frac{\sin x}{\sqrt{\pi}}, \frac{\cos 2x}{\sqrt{\pi}}, \frac{\sin 2x}{\sqrt{\pi}}, \dots$$

Entonces

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx),$$

donde

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) dx, \quad a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx,$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx.$$

Hemos obtenido así las fórmulas para el cálculo de los coeficientes del desarrollo de una función en serie trigonométrica, caso, desde luego, de que tal desarrollo sea posible ⁷.

Hemos establecido la forma de los coeficientes del desarrollo (18)

⁷ Sobre series trigonométricas ver también capítulo 12, §7.

de una función $f(x)$ mediante un sistema ortogonal de funciones en el supuesto de que dicho desarrollo exista. Sin embargo, un sistema ortogonal de funciones infinito $\phi_1, \phi_2, \dots, \phi_n, \dots$ puede resultar insuficiente para que toda función de un espacio de Hilbert posea un tal desarrollo. Para que esto sea posible, el sistema de funciones ortogonales debe satisfacer una condición adicional, la llamada condición de completitud.

Un sistema ortogonal de funciones se dice *completo* si no es posible añadirle ninguna función, no idénticamente igual a cero, que sea ortogonal a todas las funciones del sistema.

Es fácil dar ejemplos de un sistema ortogonal incompleto. Para ello tomamos un sistema ortogonal arbitrario, por ejemplo, el de las funciones trigonométricas, y suprimimos una de las funciones del sistema, por ejemplo, $\cos x$. El sistema infinito de funciones

$$1, \text{sen } x, \cos 2x, \text{sen } 2x, \dots, \cos nx, \text{sen } nx, \dots$$

es ortogonal como antes, pero desde luego no es completo, ya que la función $\cos x$ que hemos excluido es ortogonal a todas las funciones del sistema.

Si un sistema de funciones es incompleto, entonces no todas las funciones de un espacio de Hilbert se pueden desarrollar mediante él. Para comprobarlo basta intentar desarrollar mediante dicho sistema una función $f_0(x)$ distinta de cero y ortogonal a todas las funciones del sistema; entonces, por (18), todos los coeficientes deben ser cero, mientras que la función $f_0(x)$ no es idénticamente nula.

Se verifica el siguiente teorema: dado un sistema ortogonal completo de funciones en un espacio de Hilbert, $\phi_1(x), \phi_2(x), \dots, \phi_n(x), \dots$, toda función $f(x)$ se puede desarrollar en serie mediante funciones del sistema ⁸

$$f(x) = a_1\phi_1(x) + a_2\phi_2(x) + \dots + a_n\phi_n(x) + \dots$$

Los coeficientes a_n del desarrollo son iguales a las proyecciones de los vectores f sobre los elementos del sistema ortogonal

$$a_n = (f, \phi_n) = \int_a^b f(x) \phi_n(x) dx.$$

⁸ Esta serie está relacionada con su suma en el sentido definido en la fórmula (16).

El teorema de Pitágoras en un espacio de Hilbert, que fue establecido en §2, permite encontrar una interesante relación entre los coeficientes a_k y la función $f(x)$. Denotamos por $r_n(x)$ la diferencia entre $f(x)$ y la suma de los primeros n términos de su serie; es decir,

$$r_n(x) = f(x) - [a_1\phi_1(x) + \cdots + a_n\phi_n(x)].$$

La función $r_n(x)$ es ortogonal a $\phi_1(x), \phi_2(x), \cdots, \phi_n(x)$. Verifiquemos, por ejemplo, que es ortogonal a $\phi_1(x)$, es decir, que

$$\int_a^b r_n(x) \phi_1(x) dx = 0.$$

Tenemos

$$\begin{aligned} \int_a^b r_n(x) \phi_1(x) dx &= \int_a^b [f(x) - a_1\phi_1(x) - a_2\phi_2(x) - \cdots - \\ &\quad - a_n\phi_n(x)] \phi_1(x) dx = \int_a^b f(x) \phi_1(x) dx - a_1 \int_a^b \phi_1^2(x) dx^9. \end{aligned}$$

Puesto que $a_1 = \int_a^b f(x) \phi_1(x) dx$, y $\int_a^b \phi_1^2(x) dx = 1$, se deduce que $\int_a^b r_n(x) \phi_1(x) dx = 0$.

Así, en la ecuación

$$f(x) = a_1\phi_1(x) + a_2\phi_2(x) + \cdots + a_n\phi_n(x) + r_n(x) \quad (19)$$

los términos del segundo miembro son ortogonales dos a dos. Por tanto, según el teorema de Pitágoras formulado en §1, el cuadrado de la longitud de $f(x)$ es igual a la suma de los cuadrados de las longitudes de los sumandos del segundo miembro de (19); es decir,

$$\int_a^b f^2(x) dx = \int_a^b [a_1\phi_1(x)]^2 dx + \cdots + \int_a^b [a_n\phi_n(x)]^2 dx + \int_a^b r_n^2(x) dx.$$

Puesto que el sistema de funciones $\phi_1, \phi_2, \cdots, \phi_n$ está normalizado [ecuación (14)], tenemos

$$\int_a^b f^2(x) dx = a_1^2 + a_2^2 + \cdots + a_n^2 + \int_a^b r_n^2(x) dx. \quad (20)$$

* Las restantes integrales son iguales a cero, ya que las funciones $\phi_k(x)$ son ortogonales entre sí.

La serie $\sum_{k=1}^{\infty} a_k \phi_k(x)$ converge en media. Esto significa que

$$\int_a^b [f(x) - a_1 \phi_1(x) - \cdots - a_n \phi_n(x)]^2 dx \rightarrow 0,$$

es decir, que

$$\int_a^b r_n^2(x) dx \rightarrow 0.$$

Pero entonces, de la fórmula (20) obtenemos la ecuación

$$\sum_{k=1}^{\infty} a_k^2 = \int_a^b f^2(x) dx,^{10} \quad (21)$$

la cual afirma que la integral del cuadrado de una función es igual a la suma de los cuadrados de los coeficientes de su desarrollo en un sistema ortonormal cerrado de funciones. Si la condición (21) se verifica para una función arbitraria del espacio de Hilbert, recibe el nombre de condición de completitud.

Queremos llamar la atención sobre una cuestión importante. ¿Qué números a_k pueden ser coeficientes del desarrollo de una función de un espacio de Hilbert? La ecuación (21) afirma que para ello la serie $\sum_{k=1}^{\infty} a_k^2$ debe converger. Resulta además que esta condición es también suficiente; es decir, una sucesión de números a_k es la sucesión de coeficientes del desarrollo en un sistema ortogonal de una función de un espacio de Hilbert si, y sólo si, la serie $\sum_{k=1}^{\infty} a_k^2$ converge.

Observemos que este teorema fundamental se verifica si el espacio de Hilbert se interpreta como la colección de todas las funciones de cuadrado integrable en el sentido de Lebesgue (ver §2). Si nos limitáramos, por ejemplo, a las funciones continuas, la solución del problema de qué números a_k pueden ser coeficientes de un desarrollo resultaría innecesariamente complicada.

Los argumentos expuestos son sólo una de las razones que conducen al uso de la integral en un sentido generalizado (Lebesgue) en la definición del espacio de Hilbert.

¹⁰ Geométricamente, esto significa que el cuadrado de la longitud de un vector en un espacio de Hilbert es igual a la suma de los cuadrados de sus proyecciones sobre un sistema completo de direcciones ortogonales entre sí.

§4. Ecuaciones integrales

En esta sección el lector establecerá contacto con una de las más importantes e, históricamente, una de las primeras ramas del análisis funcional, a saber, la teoría de ecuaciones integrales, que ha jugado también un papel esencial en el posterior desarrollo del análisis funcional. Aparte de las exigencias internas de las matemáticas (por ejemplo, los problemas de contorno en ecuaciones diferenciales en derivadas parciales), diversos problemas de física fueron de gran importancia en el desarrollo de la teoría de ecuaciones integrales. Junto a las ecuaciones diferenciales, las ecuaciones integrales son en el siglo xx uno de los métodos matemáticos más importantes en la investigación de diversos problemas de física. En esta sección daremos algunas nociones relativas a la teoría de ecuaciones integrales. Los hechos que expondremos están íntimamente ligados con el estudio de pequeñas oscilaciones de sistemas elásticos y han surgido esencialmente (directa o indirectamente) en relación con él.

El problema de pequeñas oscilaciones de sistemas elásticos

Volvamos al problema de pequeñas oscilaciones discutidas en §2. Tratemos de encontrar las ecuaciones que describen tales oscilaciones.

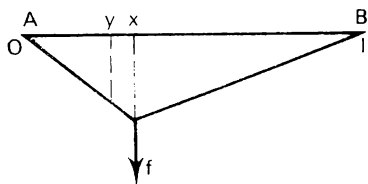


FIG. 8

Para mayor sencillez supondremos que se trata de oscilaciones de un sistema elástico lineal. Como ejemplo de tales sistemas podemos tomar una cuerda de longitud l (Fig. 8) o una barra elástica (Fig. 9). Supondremos que en la posición de equilibrio nuestro sistema elástico está situado a lo largo del segmento Ol del eje x . Aplicamos

una fuerza unitaria en el punto x . Bajo la acción de esta fuerza todos los puntos del sistema sufren un cierto desplazamiento. El desplazamiento del punto y (Fig. 8) lo denotamos por $k(x, y)$.

La función $k(x, y)$ es una función de dos puntos: el punto x en el que se aplica la fuerza, y el punto y en el que se mide el desplazamiento. Recibe el nombre de función de influencia (función de Green).

De la ley de conservación de la energía podemos deducir una importante propiedad de la función de Green $k(x, y)$; se trata de la llamada ley de reciprocidad: el desplazamiento producido en el punto y bajo la acción de una fuerza aplicada en el punto x es igual al producido en el punto x bajo la acción de la misma fuerza aplicada en el punto y . Dicho con otras palabras esto significa que

$$k(x, y) = k(y, x). \quad (22)$$

Tratemos de encontrar, por ejemplo, la función de Green para las oscilaciones longitudinales de una barra elástica (en la figura 8 hemos ilustrado desplazamientos transversales). Consideremos una barra de longitud AB , fija en los extremos (Fig. 9). En el punto C aplicamos una

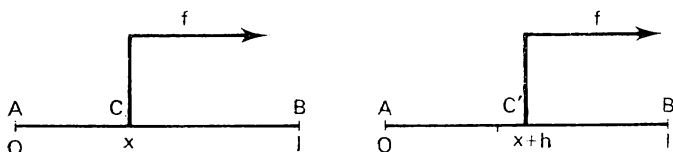


FIG. 9

fuerza que actúa en la dirección de B . Bajo la acción de esta fuerza, la barra se deforma y el punto C se desplaza a la posición C' . Denotemos la magnitud de este desplazamiento por h . Hallemos el valor de h . Mediante el valor de h podemos hallar el desplazamiento de un punto arbitrario. Para ello haremos uso de la ley de Hooke, que afirma que la fuerza es proporcional a la extensión relativa (es decir, a la razón entre el desplazamiento y la longitud). Para las compresiones se cumple una relación similar.

Bajo la acción de la fuerza f la parte AC del muelle se estira. Denotemos por T_1 la reacción producida. Al mismo tiempo, la parte CB se comprime, dando lugar a una reacción T_2 . Por la ley de Hooke tenemos:

$$T_1 = \propto \frac{h}{x}, \quad T_2 = \propto \frac{h}{l-x},$$

donde κ es el coeficiente de proporcionalidad que caracteriza las propiedades elásticas de la barra. La posición de equilibrio de las fuerzas que actúan en el punto C nos da

$$f = \kappa \frac{h}{x} + \kappa \frac{h}{l-x}; \text{ es decir, } f = \frac{\kappa h l}{x(l-x)}.$$

de donde

$$h = \frac{f}{\kappa l} x(l-x).$$

Para hallar el desplazamiento producido en un cierto punto y del segmento AC , es decir, para $y < x$, observemos que, según la ley de Hooke, para una extensión de la barra, la extensión relativa (es decir, la razón entre el desplazamiento del punto y su distancia al extremo) no depende de la posición del punto. Denotemos el desplazamiento del punto y por k .

Entonces, comparando los desplazamientos relativos de los puntos x e y obtenemos

$$\frac{k}{y} = \frac{h}{x};$$

y por tanto

$$k = h \frac{y}{x} = \frac{f}{\kappa l} y(l-x) \quad \text{para } y < x.$$

Análogamente, si el punto pertenece al segmento CB ($y > x$), obtenemos

$$k = h \frac{l-y}{l-x} = \frac{f}{\kappa l} x(l-y).$$

Teniendo en cuenta que la función de Green $k(x, y)$ es el desplazamiento del punto y bajo la acción de una fuerza unidad en el punto x ,

vemos que para las oscilaciones longitudinales de una barra elástica, la función de Green tiene la forma

$$k(x, y) = \begin{cases} \frac{1}{\kappa_l} y(l-x) & \text{para } y < x. \\ \frac{1}{\kappa_l} x(l-y) & \text{para } y > x. \end{cases}$$

De forma más o menos similar podríamos hallar la función de Green para una cuerda. Si la tensión de la cuerda es T y la longitud l , bajo la acción de una fuerza unidad aplicada en el punto x , la cuerda toma la forma ilustrada en la figura 8 y el desplazamiento en el punto y viene dado por la fórmula

$$k(x, y) = \begin{cases} \frac{1}{Tl} x(l-y) & \text{para } y > x. \\ \frac{1}{Tl} y(l-x) & \text{para } y < x, \end{cases}$$

que coincide con la función de Green que acabamos de deducir para el caso de la barra elástica.

En términos de la función de Green se puede expresar también el desplazamiento del sistema, a partir de su posición de equilibrio, producido por una fuerza distribuida continuamente de densidad $f(y)$. Puesto que sobre un intervalo de longitud Δy actúa una fuerza $f(y) \Delta y$, que se puede considerar aproximadamente concentrada en el punto y , el punto x sufre, bajo la acción de esta fuerza, un desplazamiento $k(x, y) f(y) \Delta y$. El desplazamiento bajo la acción de toda la carga es aproximadamente igual a la suma

$$\sum k(x, y) f(y) \Delta y.$$

Pasando al límite cuando $\Delta y \rightarrow 0$ vemos que el desplazamiento $u(x)$ del punto x bajo la acción de la fuerza $f(y)$ distribuida a lo largo del sistema viene dado por la fórmula

$$u(x) = \int_a^b k(x, y) f(y) dy. \quad (23)$$

Supongamos que nuestro sistema elástico no está sometido a la acción de fuerzas externas. Cuando se le desplaza de su posición de equilibrio, comienza a moverse. Estos movimientos reciben el nombre de oscilaciones libres del sistema.

Tratemos de escribir en términos de la función de Green $k(x, y)$ la ecuación que deben cumplir las oscilaciones libres del sistema elástico en cuestión. Para ello denotamos por $u(x, t)$ el desplazamiento respecto de la posición de equilibrio del punto x en el instante t . Entonces la aceleración de x en el tiempo t es igual a $\partial^2 u(x, t)/\partial t^2$.

Si ρ es la densidad lineal del sistema, es decir, ρdy la masa del elemento de longitud dy , obtenemos entonces, por una ley fundamental de la mecánica, la ecuación del movimiento sustituyendo en (23) la fuerza $f(y) dy$ por el producto de la masa por la aceleración $[\partial^2 u(y, t)/\partial t^2] \rho dy$ tomado con signo opuesto.

Así, la ecuación de las oscilaciones libres toma la forma

$$u(x, t) = - \int_a^b k(x, y) \frac{\partial^2 u(y, t)}{\partial t^2} \rho dy.$$

En la teoría de oscilaciones juegan un importante papel las llamadas oscilaciones armónicas de un sistema elástico, es decir, los movimientos para los cuales

$$u(x, t) = u(x) \text{ sen } \omega t.$$

Se caracterizan porque todo punto fijo efectúa oscilaciones armónicas (se mueve de acuerdo con una ley sinusoidal) con una cierta frecuencia ω que es la misma para todos los puntos x .

Más adelante veremos que toda oscilación libre se compone de oscilaciones armónicas.

Pongamos

$$u(x, t) = u(x) \text{ sen } \omega t$$

en la ecuación de las oscilaciones libres y eliminemos $\text{sen } \omega t$. Obtendremos entonces la siguiente ecuación que determina la función $u(x)$:

$$u(x) = \rho \omega^2 \int_a^b k(x, y) u(y) dy. \quad (24)$$

Una ecuación de este tipo recibe el nombre de ecuación integral homogénea para la función $u(x)$.

Evidentemente la ecuación (24) tiene para todo ω la solución

trivial $u(x) \equiv 0$, que corresponde al estado de reposo. Todos los valores de ω para los que existen otras soluciones de (24) distintas de cero se llaman autofrecuencias o frecuencias propias del sistema.

Puesto que no para todo valor de ω existen soluciones distintas de cero, el sistema no puede efectuar oscilaciones libres más que con ciertas frecuencias definidas. La más pequeña de ellas se llama tono fundamental del sistema, y las restantes, armónicos.

Resulta ahora que para todo sistema existe una sucesión infinita de autofrecuencia, denominada espectro de frecuencias

$$\omega_1, \omega_2, \dots, \omega_n, \dots$$

La solución $u_n(x)$ distinta de cero, correspondiente a la autofrecuencia ω_n , en la ecuación (24) nos da la forma de la oscilación característica correspondiente.

Por ejemplo, si el sistema elástico es una cuerda tensa entre los extremos O y l y sujeta a ellos, las posibles frecuencias de las oscilaciones características del sistema son

$$a \frac{\pi}{l}, 2a \frac{\pi}{l}, 3a \frac{\pi}{l}, \dots, na \frac{\pi}{l} \dots$$

donde a es un coeficiente que depende de la densidad y la tensión de

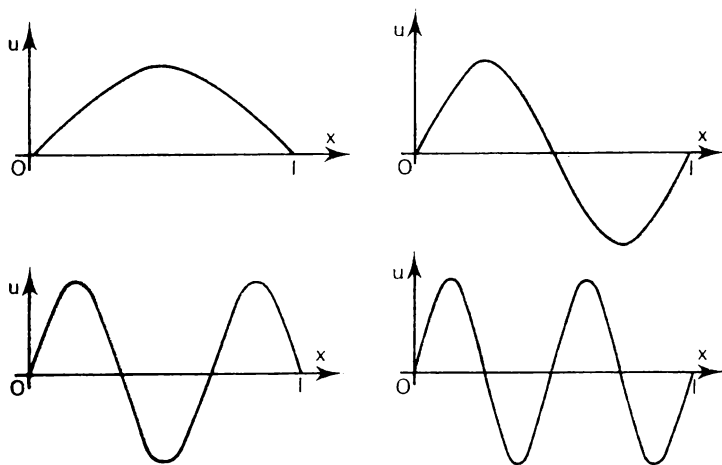


FIG. 10

la cuerda, a saber, $a = \sqrt{T/\rho}$. El tono fundamental es aquí $\omega_1 = a(\pi/l)$ y los armónicos $\omega_2 = 2\omega_1$, $\omega_3 = 3\omega_1$, $\dots$, $\omega_n = n\omega_1$. La forma de las oscilaciones armónicas correspondientes viene dada por la ecuación

$$u_n(x) = \sin \frac{n\pi}{l} x$$

y están ilustradas para $n = 1, 2, 3, 4$ en la figura 10.

Hasta aquí hemos estudiado oscilaciones libres de un sistema elástico. Si sobre el sistema elástico actúa ahora durante el movimiento una fuerza armónica, al determinar las oscilaciones armónicas bajo la acción de esta fuerza llegamos en la determinación de la función $u(x)$ a la llamada ecuación integral no homogénea

$$u(x) = \rho \omega^2 \int_a^b k(x, y) u(y) dy + h(x). \quad (25)$$

Propiedades de las ecuaciones integrales

Hasta aquí hemos conocido dos tipos de ecuaciones integrales

$$f(x) = \lambda \int_a^b k(x, y) f(y) dy \quad (26)$$

y

$$f(x) = \lambda \int_a^b k(x, y) f(y) dy + h(x), \quad (27)$$

la primera de las cuales fue obtenida en la solución del problema de las oscilaciones libres de un sistema elástico, y la segunda en la discusión de las oscilaciones sometidas a fuerzas externas.

La función incógnita en estas ecuaciones es $f(x)$. La función dada $k(x, y)$ se llama *núcleo* de la ecuación integral. La ecuación (27) recibe el nombre de *ecuación integral no homogénea*, y la 26, *ecuación lineal homogénea*. Esta se obtiene de la no homogénea cuando $h(x) = 0$.

Es claro que la ecuación homogénea tiene siempre la solución trivial, es decir, la solución $f(x) = 0$. Entre las soluciones de las dos ecuaciones existe una íntima relación.

Mencionamos a título de ejemplo el siguiente teorema: si la ecuación integral homogénea tiene solamente la solución cero, entonces la correspondiente ecuación no homogénea tiene solución para toda función $h(x)$.

Si para un cierto valor λ una ecuación homogénea tiene la solución $f(x)$, no idéntica a cero, entonces este valor λ se dice que es un *autovalor*, y su correspondiente solución $f(x)$, una *autofunción*. Hemos visto ya que, cuando una ecuación integral describe las oscilaciones libres de un sistema elástico, los autovalores están íntimamente ligados con las frecuencias de las oscilaciones del sistema (esto es, $\lambda = \rho\omega^2$). Las autofunciones dan entonces la forma de las correspondientes oscilaciones armónicas.

En los problemas sobre oscilaciones se deduce de la ley de la conservación de la energía que

$$k(x, y) = k(y, x). \quad (28)$$

Un núcleo que satisfaga la condición (28) se llama *simétrico*.

Las autofunciones y autovalores de una ecuación con un núcleo simétrico tiene importantes propiedades. Se puede probar que una ecuación tal tiene siempre una sucesión de autovalores reales

$$\lambda_1, \lambda_2, \dots, \lambda_n, \dots$$

A cada autovalor corresponde una o varias autofunciones. Las autofunciones correspondientes a autovalores distintos son siempre ortogonales entre sí¹¹.

Así, pues, para toda ecuación integral con un núcleo simétrico el sistema de autofunciones es un sistema ortogonal de funciones. Se plantea entonces la cuestión de saber cuándo este sistema es completo, es decir, cuándo se puede expresar toda función del espacio de Hilbert en forma de serie mediante un sistema de autofunciones de la ecuación integral. En particular, si la ecuación

$$\int_a^b k(x, y) f(y) dy = 0 \quad (29)$$

¹¹ Este teorema se probará en la próxima sección.

se satisface solamente para $f(y) \equiv 0$, el sistema de autofunciones de la ecuación integral

$$\lambda \int_a^b k(x, y) f(y) dy = f(x)$$

es un sistema ortogonal completo¹².

En consecuencia, toda función $f(x)$ de cuadrado integrable puede expresarse en este caso como un desarrollo en serie de autofunciones. Al discutir los diversos tipos de ecuaciones integrales, obtenemos un método general y potente para probar que existen diversos sistemas ortogonales que son cerrados, es decir, que las funciones se pueden desarrollar en serie de funciones ortogonales. Por este método podemos probar la completitud del sistema de funciones trigonométricas, de funciones cilíndricas, esféricas, y muchos otros sistemas importantes de funciones.

El hecho de que una función arbitraria se pueda desarrollar en serie de autofunciones quiere decir, en el caso de las oscilaciones, que toda oscilación se puede descomponer en una suma de oscilaciones armónicas. Tal descomposición da lugar a un método aplicable a la resolución de problemas sobre oscilaciones en varios dominios de la mecánica y la física (oscilaciones de los cuerpos elásticos, oscilaciones acústicas, ondas electromagnéticas, etc.).

El desarrollo de la teoría de ecuaciones integrales lineales impulsó la creación de la teoría general de operadores lineales, de la que la teoría de ecuaciones integrales lineales constituye una parte orgánica. En las últimas décadas, los métodos generales de la teoría de operadores lineales ha contribuido vigorosamente al desarrollo de la teoría de ecuaciones integrales.

§5. Operadores lineales y desarrollos ulteriores del análisis funcional

En la sección precedente hemos visto que los problemas sobre las oscilaciones de un sistema elástico conducen a la investigación de autovalores y autofunciones de ecuaciones integrales. Observemos que estos problemas pueden reducirse también a la investigación de

¹² En el caso en que $k(x, y)$ es la función de Green de un sistema elástico, la ecuación (29) adquiere un significado físico sencillo. En efecto [ver fórmula (23)], hemos visto que bajo la acción de una fuerza $f(y)$ distribuida a lo largo del sistema, el desplazamiento del sistema a partir de su posición de equilibrio se expresa por la fórmula $u(x) = \int_a^b k(x, y) f(y) dy$. Así, pues, la condición (29) significa que toda fuerza distinta de cero desplaza el sistema de su posición de equilibrio.

autovalores y autofunciones de ecuaciones diferenciales lineales¹³. Muchos otros problemas físicos conducen también al problema de calcular los autovalores y autofunciones de ecuaciones diferenciales o integrales lineales.

Demos un ejemplo más. En radiotecnia moderna las llamadas guías de onda se utilizan ampliamente para la transmisión de oscilaciones electromagnéticas de alta frecuencia; las guías de onda vienen a ser tubos metálicos huecos en los que se propagan las ondas electromagnéticas. Se sabe que en una guía de onda sólo pueden propagarse oscilaciones electromagnéticas de longitud de onda no demasiado grande. El problema de hallar la longitud de onda crítica conduce al problema de los autovalores de una cierta ecuación diferencial.

Aparecen además problemas sobre autovalores en álgebra lineal, en la teoría de ecuaciones diferenciales ordinarias, en cuestiones de estabilidad, etc.

De este modo, llegó un momento en que fue necesario discutir todos estos problemas afines desde un punto de vista único. Este punto de vista común es la teoría general de operadores lineales. Muchos de los problemas sobre autovalores y autofunciones que se plantean en diversos casos concretos, sólo llegaron a entenderse a la luz de la teoría general de operadores. En consecuencia, en esta y otras muchas direcciones la teoría general de operadores resultó ser una herramienta de investigación muy fructífera en aquellos dominios de la matemática en que es aplicable.

En el posterior desarrollo de la teoría de operadores desempeñó un papel muy importante la mecánica cuántica, puesto que hace uso extensivo de los métodos de esta teoría. El aparato matemático fundamental de la mecánica cuántica es la teoría de los llamados operadores autoadjuntos. La formulación de los problemas matemáticos que surgen en la mecánica cuántica fue, y es todavía, un poderoso estímulo para el desarrollo del análisis funcional.

El método de operadores en ecuaciones diferenciales o integrales demostró ser también extremadamente útil para el desarrollo de métodos prácticos de aproximación de soluciones en tales ecuaciones.

Conceptos fundamentales de la teoría de operadores

Pasemos ahora a dar una explicación de las definiciones y hechos fundamentales de la teoría de operadores.

En análisis hemos encontrado el concepto de función. En su

¹³ Ver capítulo 6, §5.

forma más simple ésta era una relación que asocia a cada número x (el valor de la variable independiente) un número y (el valor de la función). En el desarrollo posterior del análisis se hace necesario considerar relaciones de un tipo más general.

Tales relaciones más generales se discuten, por ejemplo, en el cálculo de variaciones (capítulo 8), donde a cada función se le asocia un número. Si a toda función se le asocia un cierto número decimos entonces que se ha definido un funcional. Como ejemplo de funcional podemos tomar la correspondencia entre una función arbitraria $y = f(x)$ ($a \leq x \leq b$) y la longitud de arco de la curva representada por ella. Obtenemos otro ejemplo de funcional si asociamos a cada función $y = f(x)$ ($a \leq x \leq b$) su integral definida $\int_a^b f(x) dx$.

Si consideramos $f(x)$ como un punto de un espacio de dimensión infinita, entonces un funcional es simplemente una función de los puntos de dicho espacio de dimensión infinita. Desde este punto de vista, los problemas del cálculo de variaciones se refieren a la búsqueda de máximos y mínimos de funciones de los puntos del espacio de dimensión infinita.

Para aclarar lo que entendemos por funcionales continuos es necesario definir primero lo que entendemos por proximidad de dos puntos de un espacio de dimensión infinita. En la sección 2 dimos la distancia entre dos funciones $f(x)$ y $g(x)$ (puntos de un espacio de dimensión infinita) como

$$\sqrt{\int_a^b [f(x) - g(x)]^2 dx}.$$

Este método de asignar una distancia a un espacio de dimensión infinita es de uso frecuente, pero por supuesto no es el único posible. En otros problemas pueden ser más adecuados otros métodos de dar la distancia entre funciones. Podemos señalar, por ejemplo, el problema de la teoría de aproximación de funciones (ver capítulo 12, §3) donde la distancia entre funciones —que caracteriza la medida de la proximidad de las dos funciones $f(x)$ y $g(x)$ — se da, por ejemplo, mediante la fórmula

$$\max |f(x) - g(x)|.$$

Otros métodos de dar una distancia entre funciones aparecen en el estudio de funcionales en el cálculo de variaciones. Métodos dis-

tintos de definir la distancia entre funciones originan espacios distintos de dimensión infinita.

De este modo, los distintos espacios (funcionales) de dimensión infinita difieren unos de otros por el conjunto de funciones y por la definición de distancia entre ellas. Por ejemplo, si tomamos el conjunto de funciones de cuadrado integrable y definimos la distancia como

$$\int_a^b [f(x) - g(x)]^2 dx,$$

llegamos entonces al espacio de Hilbert que se introdujo en el §2; pero si tomamos el conjunto de las funciones continuas y definimos las distancias como $\max |f(x) - g(x)|$, obtenemos entonces lo que se llama el espacio (C).

Al discutir las ecuaciones integrales encontramos expresiones de la forma

$$g(x) = \int_a^b k(x, y) f(y) dy.$$

Para un núcleo dado $k(x, y)$ esta ecuación define una regla mediante la cual se pone cada función $f(x)$ en correspondencia con otra función $g(x)$.

Este tipo de correspondencia que asocia a cada función f otra función g se llama un *operador*.

Diremos que se ha definido un operador lineal A en un espacio de Hilbert si disponemos una regla que asocie a cada función f otra función g . La correspondencia no necesita estar definida para todas las funciones del espacio de Hilbert. En este caso el conjunto de las funciones f para las que existe la función $g = Af$ se llama *dominio de definición* del operador A (semejante al dominio de definición en el análisis ordinario). La correspondencia misma se denota usualmente como sigue:

$$g = Af. \quad (30)$$

El carácter lineal del operador significa que la suma de las funciones f_1, f_2 va asociada con la suma de Af_1 y Af_2 , y que el producto de f por un número λ va asociado con la función λAf ; esto es

$$A(f_1 + f_2) = Af_1 + Af_2 \quad (31)$$

$$y \quad A(\lambda f) = \lambda Af. \quad (32)$$

En ocasiones se postula también la continuidad para operadores lineales; es decir, se exige que la convergencia de una sucesión de funciones f_n a una función f implique que la sucesión Af_n converja, a su vez, hacia Af .

Demos algunos ejemplos de operadores lineales.

1. Asociemos a cada función $f(x)$ la función $g(x) = \int_a^x f(t) dt$, esto es, la integral indefinida de f . La linealidad de este operador se desprende de las propiedades ordinarias de la integral, a saber, del hecho de que la integral de la suma es igual a la suma de las integrales y de que un factor constante se puede sacar siempre fuera del signo integral.

2. Asociemos a cada función diferenciable $f(x)$ su derivada $f'(x)$. Este operador se designa por la letra D ; esto es:

$$f'(x) = Df(x).$$

Observemos que este operador no está definido para todas las funciones del espacio de Hilbert, sino solamente para aquellas que tienen una derivada que pertenezca al espacio de Hilbert. Estas funciones forman, como hemos dicho previamente, el dominio de definición de este operador.

3. Los ejemplos 1 y 2 son ejemplos de operadores lineales en espacios de dimensión infinita. Pero en otros capítulos de este libro han aparecido también ejemplos de operadores en espacios de dimensión finita. Así, en el capítulo 3 se estudiaron las transformaciones afines. Si una transformación afín deja fijo el origen de coordenadas en un plano o en el espacio, tenemos entonces un ejemplo de operador lineal en un espacio bidimensional o tridimensional. Las transformaciones lineales de un espacio n -dimensional que introdujimos en el capítulo 15 aparecen ahora como operadores lineales en un espacio n -dimensional.

4. En ecuaciones integrales hemos encontrado también una clase muy importante y útil de operadores lineales: los llamados operadores integrales. Tomemos una cierta función $k(x, y)$. Entonces la fórmula

$$g(x) = \int_a^b k(x, y) f(y) dy$$

asocia a cada función f una cierta función g . Podemos escribir simbólicamente esta transformación de la siguiente forma:

$$g = Af.$$

El operador A se dice que es un operador integral. Podríamos mencionar muchos otros ejemplos importantes de operadores lineales.

En la sección 4 hablábamos de la ecuación integral no homogénea.

$$f(x) = \lambda \int_a^b k(x, y) f(y) dy + h(x).$$

Con la notación de la teoría de operadores podemos volver a escribir la fórmula de la siguiente forma:

$$f = \lambda Af + h, \quad (33)$$

donde λ es un número dado, h una función dada (un vector de un espacio de dimensión infinita), y f la función buscada. Con la misma notación, la ecuación homogénea se escribe como sigue:

$$f = \lambda Af, \quad (34)$$

Los teoremas clásicos sobre ecuaciones integrales, tal como, por ejemplo, el teorema formulado en la sección 4 sobre la relación entre la solubilidad de la ecuación integral no homogénea y la homogénea, no son ciertos para cualquier ecuación entre operadores. Sin embargo, se pueden imponer ciertas condiciones generales al operador A de forma que los teoremas sí sean válidos.

Estas condiciones se formulan en términos topológicos y expresan que el operador A debe transformar la esfera unidad (es decir, el conjunto de los vectores cuya longitud no excede de 1) en un conjunto compacto.

Autovalores y autovectores de operadores

El problema de los autovalores y autofunciones de una función integral, al que nos condujo el problema de las oscilaciones, se puede formular del siguiente modo: hallar los valores de λ para los que existe una función f distinta de 0 que satisface la ecuación

$$f(x) = \lambda \int_a^b k(x, y) f(y) dy.$$

Como antes, esta ecuación se puede escribir así:

$$f = \lambda Af$$

o

$$Af = \frac{1}{\lambda} f. \quad (35)$$

Si por A representamos ahora un operador lineal arbitrario, entonces un vector f que satisface la ecuación (35) se dice que es un autovector del operador A , y el número $1/\lambda$, el correspondiente autovalor.

Puesto que el vector $(1/\lambda)f$ coincide en dirección con el vector f (difiere de f sólo en un factor numérico), se puede decir que el problema de encontrar autovectores es el problema de hallar vectores f distintos de 0 que no cambian de dirección cuando se les aplica la transformación A .

Esta forma de contemplar los autovalores permiten unificar el problema de los autovalores de una ecuación integral (si A es un operador integral), de ecuaciones diferenciales (si A es un operador diferencial), y el problema de los autovalores en álgebra lineal (si A es una transformación lineal en espacios de dimensión finita; ver capítulo 6 y capítulo 15). En el caso de espacios tridimensionales este problema surge al hallar los llamados ejes principales de un elipsoide.

En el caso de las ecuaciones integrales, existen varias propiedades importantes (el hecho de ser reales los autovalores, la ortogonalidad de las autofunciones, etc.) que son consecuencia de la simetría del núcleo, esto es, de la ecuación $k(x, y) = k(y, x)$.

De un operador arbitrario A en un espacio de Hilbert que cumple una propiedad análoga a la anterior se dice que es un operador autoadjunto.

En general, para que un operador A sea autoadjunto se tiene que cumplir que para cada dos elementos f_1 y f_2 se verifique la ecuación

$$(Af_1, f_2) = (f_1, Af_2)$$

donde (Af_1, f_2) denota el producto escalar del vector Af_1 y el vector f_2 .

En problemas de mecánica, la condición de un operador de ser autoadjunto es en general una consecuencia de la ley de conservación

de la energía. Se satisface, por tanto, para operadores relacionados, por ejemplo, con oscilaciones para las que no existe pérdida (disipación) de energía.

La mayoría de los operadores que aparecen en mecánica cuántica son también autoadjuntos.

Probemos que un operador integral con núcleo $k(x, y)$ simétrico es autoadjunto. En efecto, en este caso Af_1 es la función $\int_a^b k(x, y) f_1(y) dy$. Por tanto, el producto escalar (Af_1, f_2) , que es igual a la integral del producto de esta función por f_2 , viene dado por la fórmula

$$(Af_1, f_2) = \int_a^b \int_a^b k(x, y) f_1(y) f_2(x) dy dx.$$

De modo similar

$$(f_1, Af_2) = \int_a^b \int_a^b k(x, y) f_2(y) f_1(x) dy dx.$$

La ecuación $(Af_1, f_2) = (f_1, Af_2)$ es consecuencia inmediata de la simetría del núcleo $k(x, y)$.

Los operadores autoadjuntos tienen una serie de importantes propiedades muy útiles en las aplicaciones de estos operadores a la solución de diversos problemas. Por ejemplo, los autovalores de un operador lineal autoadjunto son siempre reales, y las autofunciones correspondientes a autovalores distintos son ortogonales entre sí.

Probemos, por ejemplo, la última aserción. Sean λ_1 y λ_2 dos autovalores distintos del operador A , y f_1 y f_2 , los correspondientes autovectores. Esto significa que

$$\begin{aligned} Af_1 &= \lambda_1 f_1, \\ Af_2 &= \lambda_2 f_2. \end{aligned} \tag{36}$$

Efectuamos el producto escalar de la primera ecuación (36) por f_2 , y de la segunda por f_1 . Tenemos entonces

$$\begin{aligned} (Af_1, f_2) &= \lambda_1 (f_1, f_2), \\ (Af_2, f_1) &= \lambda_2 (f_2, f_1). \end{aligned} \tag{37}$$

Puesto que el operador A es autoadjunto, tenemos $(Af_1, f_2) = (Af_2, f_1)$. Si restamos la segunda ecuación (37) de la primera, obtenemos

$$0 = (\lambda_1 - \lambda_2) (f_1, f_2).$$

Como $\lambda_1 \neq \lambda_2$, tenemos $(f_1, f_2) = 0$, esto es, los autovectores f_1, f_2 son ortogonales.

El estudio de los operadores autoadjuntos ha proporcionado gran claridad a muchos problemas y cuestiones concretos relacionados con la teoría de autovalores. Profundicemos en uno de ellos: el problema del desarrollo mediante autofunciones en el caso de un espectro continuo.

Para explicar lo que significa un espectro continuo volvamos de nuevo al ejemplo clásico de las oscilaciones de una cuerda. Hemos visto ya que las frecuencias características de las oscilaciones de una cuerda de longitud l toma la sucesión de valores

$$a \frac{\pi}{l}, 2a \frac{\pi}{l}, \dots, na \frac{\pi}{l}, \dots$$

Llevemos los valores de esta sucesión sobre el eje numérico $O\lambda$. Si aumentamos la longitud l de la cuerda, la distancia entre dos puntos contiguos de la sucesión disminuirá, llenando más densamente el eje numérico. En el límite, cuando $l \rightarrow \infty$, es decir, para una cuerda infinita, las autofrecuencias llenan todo el semieje numérico $\lambda \geq 0$. En este caso se dice que el sistema tiene un espectro continuo.

Ya dijimos que para una cuerda de longitud l , el desarrollo en serie mediante autofunciones es una expresión en senos y cosenos de $n(\pi/l)x$; esto es, una serie trigonométrica

$$f(x) = \frac{a_0}{2} + \sum \left(a_n \cos n \frac{\pi}{l} x + b_n \operatorname{sen} n \frac{\pi}{l} x. \text{ II} \right).$$

De nuevo se puede probar que para el caso de una cuerda infinita, una función más o menos arbitraria se puede desarrollar mediante senos y cosenos. Sin embargo, como las autofrecuencias están ahora distribuidas a lo largo de la recta numérica, no obtenemos un desarrollo en serie, sino una integral de Fourier:

$$f(x) = \int_{-\infty}^{+\infty} [A(\lambda) \cos \lambda x + B(\lambda) \operatorname{sen} \lambda x] d\lambda.$$

El desarrollo en una integral de Fourier era ya conocido y ampliamente utilizado en el siglo XIX en la solución de diversos problemas de la física matemática.

Sin embargo, en casos más generales de espectro continuo¹⁴, muchos de los problemas referentes al desarrollo de funciones mediante autofunciones no estaban suficientemente claros. Sólo la creación de la teoría general de operadores autoadjuntos proporcionó la claridad necesaria a estos problemas.

Mencionemos todavía otra serie de problemas clásicos que han sido resueltos a partir de la teoría general de operadores. La discusión de oscilaciones en las que hay pérdida (dispersión) de energía pertenece a tales problemas.

En este caso podemos buscar de nuevo las oscilaciones libres del sistema de la forma $u(x)\phi(t)$. Sin embargo, en contraposición al caso de las oscilaciones sin pérdida de energía, la función $\phi(t)$ no es simplemente $\cos \omega t$, sino que tiene la forma $e^{-kt} \cos \omega t$, donde $k > 0$. De este modo, la solución correspondiente tiene la forma $u(x)e^{-kt} \cos \omega t$. En este caso cada punto x efectúa de nuevo oscilaciones (con frecuencia ω), pero estas oscilaciones son amortiguadas, ya que para $t \rightarrow \infty$ la amplitud de las mismas tiende a 0 debido al factor e^{-kt} .

Es conveniente escribir las oscilaciones características del sistema en la forma compleja $u(x)e^{-i\lambda t}$, donde, si no hay rozamiento, el número λ es real, y si lo hay, λ es complejo.

El problema de las oscilaciones de un sistema con disipación de energía conduce una vez más a un problema de autovalores, pero esta vez no para operadores autoadjuntos. Un hecho característico es la presencia de autovalores complejos, indicativo del amortiguamiento de las oscilaciones libres.

Recurriendo a un método de la teoría de operadores, junto con métodos de la teoría de funciones analíticas, M. V. Keldys estudió esta clase de problemas en 1950-1951 y probó para ellos el carácter completo del sistema de autofunciones.

Relación del análisis funcional con otras ramas de las matemáticas y la mecánica cuántica

Hemos mencionado ya que la creación de la mecánica cuántica impulsó decididamente el desarrollo del análisis funcional. Del mismo modo que las necesidades de la mecánica y la física clásica determinaron el nacimiento del cálculo diferencial e integral, así también el desarrollo del análisis funcional fue, y todavía es, el resultado de la vigorosa influencia de la física contemporánea,

¹⁴ Podemos tomar como ejemplo las oscilaciones de un medio elástico no homogéneo y también muchos problemas de mecánica cuántica.

principalmente de la mecánica cuántica. El aparato matemático fundamental de la mecánica cuántica consta de las ramas de las matemáticas relacionadas esencialmente con el análisis funcional. Sólo podemos indicar brevemente las relaciones que prevalecen aquí, ya que una exposición del desarrollo de los fundamentos de la mecánica cuántica rebasaría el marco de este libro.

En mecánica cuántica, la descripción matemática del estado del sistema se hace mediante un espacio de Hilbert. Magnitudes tales como la energía, el impulso y la cantidad de movimiento se estudian por medio de operadores autoadjuntos. Por ejemplo, los posibles niveles de energía de un electrón en un átomo se calculan como autovalores del operador de energía. Las diferencias entre estos autovalores dan las frecuencias de los cuantos de luz emitidos, definiendo así la estructura del espectro de radiación de la sustancia dada. Los correspondientes estados del electrón se describen como autofunciones del operador de energía.

La solución de los problemas de la mecánica cuántica requiere a menudo el cálculo de autovalores de distintos operadores (por lo general diferenciales). En algunos casos complicados la solución exacta de estos problemas resulta prácticamente imposible. La llamada teoría de perturbaciones se utiliza ampliamente para hallar una solución aproximada de estos problemas; esta teoría permite calcular, a partir de los autovalores y autofunciones de un cierto operador autoadjunto A , los autovalores de otro operador A_1 que difiere muy poco de aquél. Es de observar que la teoría de perturbaciones no cuenta aún con una fundamentación matemática completa, lo cual constituye un interesante e importante problema matemático.

Independientemente de la determinación aproximada de autovalores, a veces es posible saber bastante acerca de un problema por medio de la investigación cualitativa. Estos estudios proceden en problemas de mecánica cuántica sobre la base de las simetrías existentes en cada caso. Como ejemplos de estas simetrías podemos tomar las propiedades de simetría de los cristales, simetría esférica en un átomo, simetría respecto a un eje de rotación, y otros. Como las simetrías forman grupo (ver capítulo 19), los métodos de la teoría de grupos (la llamada teoría de representación de grupos) permiten resolver un cierto número de problemas sin efectuar ningún cálculo. Como ejemplos debemos mencionar la clasificación de los espectros atómicos, las transformaciones nucleares y otros problemas. Vemos así cómo la mecánica cuántica hace uso extensivo del aparato matemático de la teoría de los operadores autoadjuntos. Al mismo tiempo,

el desarrollo incesante de la mecánica cuántica conduce a un desarrollo continuado de la teoría de operadores al plantearle a ésta nuevos problemas.

La influencia de la mecánica cuántica y el propio desarrollo matemático interno del análisis funcional han determinado que en los últimos años los problemas y métodos algebraicos desempeñaran un importante papel en el análisis funcional. Esta intensificación de las tendencias algebraicas en el análisis contemporáneo puede equipararse a la progresiva importancia de los métodos algebraicos en la física teórica contemporánea en comparación con los métodos de la física del siglo XIX.

En conclusión, deseamos subrayar una vez más que el análisis funcional es una de las ramas de más rápido desarrollo de las matemáticas contemporáneas. Sus relaciones y aplicaciones en la física contemporánea, en ecuaciones diferenciales, en cálculos aproximados y el uso que hace de métodos generales desarrollados en álgebra, topología, teoría de funciones de una variable real, etc., hacen del análisis funcional uno de los focos de la matemática actual.

Capítulo 19

GRUPOS Y OTROS SISTEMAS ALGEBRAICOS

§1. *Introducción*

En el capítulo 4, que trataba del álgebra de polinomios, hemos hablado ya de las principales líneas de desarrollo del álgebra, de su lugar entre las demás disciplinas matemáticas y de los cambios en cuanto al objeto de su estudio. El objeto de este capítulo es dar al lector una idea de aquellas nuevas teorías algebraicas que, aunque nacidas en el siglo pasado, no han alcanzado un desarrollo completo hasta el siglo presente, ejerciendo un profundo impacto en la investigación matemática contemporánea.

El álgebra contemporánea —y también la clásica— es el estudio de las operaciones, de las reglas de cálculo. Pero no se circunscribe al estudio de las propiedades de las operaciones con números, sino que aspira a investigar propiedades de operaciones con elementos de una naturaleza mucho más general. Esta tendencia viene dictada por necesidades de orden práctico. Por ejemplo, en mecánica sumamos fuerzas, velocidades, rotaciones. En el álgebra lineal (ver capítulo 15), cuyas ideas y métodos encuentran gran aplicación en los cálculos prácticos, los dominios de las operaciones son matrices, transformaciones lineales o vectores de un espacio n -dimensional.

La teoría de grupos desempeña un papel particularmente importante en el álgebra contemporánea; gran parte de este capítulo está

dedicado a ella. Entre otras teorías algebraicas, trataremos de la teoría de sistemas hipercomplejos, que constituye un necesario e importante paso en el proceso histórico del desarrollo del concepto de número. Desde luego, estas dos teorías no agotan ni mucho menos el contenido del álgebra actual, pero ilustran bastante bien sus ideas y métodos.

La teoría de grupos ha nacido de la necesidad de encontrar un método para estudiar propiedades tan importantes del mundo real como es, por ejemplo, la simetría.

El conocimiento de las propiedades simétricas de los cuerpos geométricos u otros objetos matemáticos o físicos nos da a veces la clave para determinar su estructura. Sin embargo, aunque el concepto de simetría es completamente intuitivo, para dar una descripción exacta y general de lo que es una simetría, y en particular una exposición cuantitativa de sus propiedades, es necesario hacer uso del aparato de la teoría de grupos.

La teoría de grupos nació hace bastante tiempo, a finales del siglo XVIII y comienzos del XIX. Al principio se desarrolló solamente como un método auxiliar en los problemas sobre la solución por medio de radicales de ecuaciones de grado superior. Ello se debió a que fue precisamente en estos problemas donde se observó por primera vez que las propiedades de equivalencia, de simetría de las raíces de la ecuación, son fundamentales para la resolución completa del problema. A lo largo de los siglos XIX y XX se puso de manifiesto en muchas otras ramas de la ciencia el importante papel de las leyes de simetría: geometría, cristalografía, física y química. Esto llevó a una mayor propagación de los métodos y resultados de la teoría de grupos. Puesto que todo campo de aplicación presentaba problemas peculiares propios a la teoría de grupos, el creciente número de estos dominios ejerció también el efecto opuesto, dando lugar a nuevas ramas de la teoría de grupos. El resultado de todo ello fue que la teoría contemporánea de grupos, que es una entidad única en sus conceptos esenciales, se divide en un cierto número de disciplinas más o menos independientes: la teoría general de grupos, la teoría de grupos finitos, la teoría de grupos continuos y de grupos discretos de transformaciones, la teoría de representaciones y caracteres de grupos, y otras. A lo largo de su gradual evolución, los métodos y conceptos de la teoría de grupos resultaron ser importantes no sólo para el estudio de las leyes de simetría, sino también para la solución de muchos otros problemas.

En nuestro tiempo, el concepto de grupo se ha convertido en

uno de los conceptos generales más importantes de la matemática moderna, y la teoría de grupos ocupa un importante lugar entre las disciplinas matemáticas. De forma sobresaliente han contribuido al desarrollo de la teoría de grupos y sus aplicaciones E. S. Fedorov, O. Ju. Smidt y L. S. Pontriagin. En los tiempos actuales las investigaciones de los matemáticos soviéticos ocupan un lugar preponderante.

§2. Simetrías y transformaciones

Las formas más sencillas de simetría

Comenzamos con una exposición de las formas de simetría más sencillas, con las que el lector está familiarizado por la experiencia cotidiana. Una de ellas es la simetría especular, o simetría con respecto a un plano.

Se dice que un punto A del plano es simétrico con el punto B respecto a un plano α (Fig. 1) si el plano corta perpendicularmente a AB en su punto medio. Se dice también que B es la imagen simétrica de A respecto al plano α . Un cuerpo geométrico se dice que es simétrico respecto a un plano si dicho plano divide al cuerpo en dos partes que son simétricas respecto a él. Se dice que este plano es un plano de simetría del cuerpo. Las simetrías especulares se encuentran a menudo en la naturaleza. Por ejemplo, el cuerpo humano, o el cuerpo de los pájaros y en general de los animales suelen tener un plano de simetría.

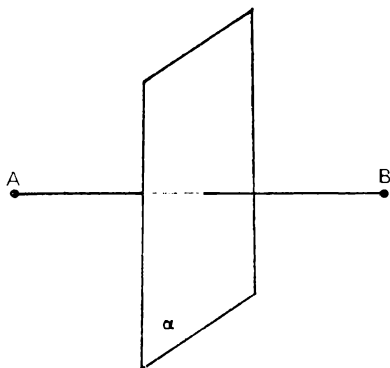


FIG. 1

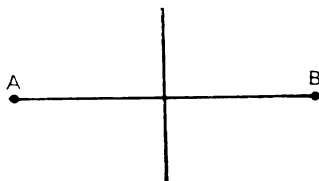


FIG. 2

La simetría respecto a una recta se define de modo similar. Se dice que los puntos A, B son simétricos respecto de una recta si dicha recta corta perpendicularmente al segmento AB en su punto medio (Fig. 2). Se dice que un cuerpo geométrico es simétrico respecto de una recta o que tiene esta recta como eje de simetría de orden 2 si para cada punto del cuerpo su simétrico pertenece también al cuerpo.

Un cuerpo que tiene un eje de simetría de orden 2 vuelve a coincidir consigo mismo cuando se le somete a una rotación de 180° alrededor de dicho eje.

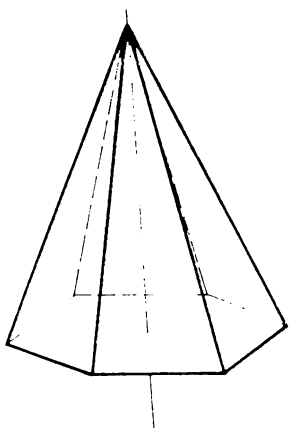


FIG. 3

El concepto de eje de simetría se puede generalizar de un modo natural. Se dice que una recta es un eje de simetría de orden n para un cuerpo dado si el cuerpo coincide consigo mismo cuando se le somete a una rotación alrededor de dicha recta de ángulo $1/n \cdot 360^\circ$. Por ejemplo, una pirámide regular cuya base es un polígono regular de n lados tiene la línea que une su vértice con el centro de la base (figura 3) como eje de simetría de orden n .

Decimos que una recta es un eje de rotación de un cuerpo si, para todo giro de eje dicha recta y ángulo arbitrario, el cuerpo coincide consigo mismo. Por

ejemplo, el eje de un cilindro, el de un cono o el diámetro de una esfera son ejes de rotación. Un eje de rotación es también un eje de simetría de cualquier orden.

Por último, un tipo importante de simetría es la simetría respecto a un punto, o simetría central. Dos puntos A y B se dice que son simétricos respecto a O si el segmento AB tiene a O como punto medio. Se dice que un cuerpo es simétrico respecto a un centro O si el simétrico de cualquier punto del cuerpo respecto a O pertenece también al cuerpo. Como ejemplo tenemos la esfera y el cubo, cuyos centros son centros de simetría (Fig. 4).

El conocimiento de los planos, ejes y centros de simetría de un cuerpo da una idea muy completa de sus propiedades de simetría.

Pero el concepto de simetría también tiene sentido cuando se aplica a entidades que no son figuras geométricas. Por ejemplo, la afirmación de que en la ecuación $x_1^3 + x_2^3 + x_3^3 + x_4^3$ las variables

x_1, x_2, x_3, x_4 aparecen simétricamente tiene un sentido perfectamente claro; también lo tiene decir que en el polinomio $x_1^3 + x_2^2 + x_3^2 + x_4^3$ las variables x_1 y x_4 , y x_2, x_3 aparecen simétricamente; en cambio, es claro que las variables x_1, x_2 juegan papeles diferentes. El número de tales ejemplos es interminable. Esto nos lleva a plantear la siguiente cuestión: ¿qué es, en general, una simetría y cómo se puede definir matemáticamente la relación de simetría?

Resulta que la contestación precisa a esta cuestión está relacionada con el concepto de transformación, que ya ha aparecido muchas veces a lo largo de este libro, incluso en los primeros capítulos. Para poder dar una definición general de simetría que abarque casos tan heterogéneos

como los de simetría de cuerpos espaciales y simetría de polinomios es necesario formular el concepto de transformación de forma muy general.

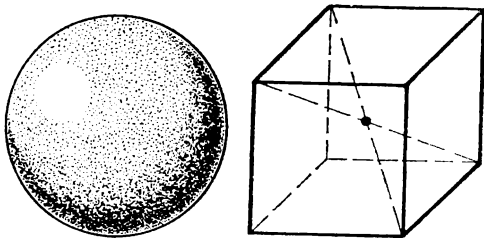


FIG. 4

Transformaciones

Sea M una colección finita o infinita de objetos completamente arbitrarios. M puede ser, por ejemplo, el conjunto de números $1, 2, \dots, n$, el conjunto de las variables independientes x_1, x_2, x_3, x_4 , o el conjunto de los puntos de un plano. Si asociamos a cada elemento de M otro elemento también de M bien definido, se dice que hemos dado una transformación de M . Toda transformación de un conjunto finito M se puede dar mediante una tabla formada por dos filas: en la superior escribimos los nombres de los elementos de M y debajo de cada uno de ellos el nombre del elemento que le corresponde en la transformación. Por ejemplo, la tabla

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 2 & 1 \end{pmatrix}$$

denota la transformación del conjunto de números $1, 2, 3, 4$ mediante la cual los números $1, 2, 3, 4$ se transforman, respectivamente, en los

números 2, 3, 2 1. Si ponemos en la fila superior los números 1, 2, 3, 4 en el orden 3, 4, 1, 2, podemos escribir la misma transformación también en la forma

$$\begin{pmatrix} 3 & 4 & 1 & 2 \\ 2 & 1 & 2 & 3 \end{pmatrix}.$$

Si el conjunto M es infinito, pero se pueden contar (enumerar) sus elementos, la transformación se puede dar de forma similar escribiendo los elementos en una única fila (por ejemplo, si M es el conjunto de los números naturales 1, 2, 3, ...).

En el estudio de las transformaciones es necesario introducir una notación adecuada. Denotaremos las transformaciones simplemente por letras A, B , etc., y si alguna de ellas, por ejemplo A , denota una transformación de M , entonces denotamos por mA (donde m es un elemento arbitrario de M) el transformado de m por A . Supongamos, por ejemplo, que

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 2 & 1 \end{pmatrix}; \text{ entonces } 1A = 2, 2A = 3, 3A = 2, 4A = 1.$$

Indiquemos algunas transformaciones que juegan un importante papel en geometría.

Consideremos una recta arbitraria a en el espacio y asociemos a cada punto P del espacio el punto Q obtenido al girar el punto P un ángulo fijo ϕ alrededor del eje a (Fig. 5). Hemos definido de esta forma una transformación del conjunto de todos los puntos del espacio; esta transformación se denomina giro de ángulo ϕ alrededor del eje a .

Observemos que la palabra «giro» en mecánica denota un cierto proceso como resultado del cual los puntos del cuerpo toman una nueva posición. Aquí hemos usado el término «giro» en el sentido de una transformación del espacio. Hacemos abstracción de la forma en que se realiza el proceso, interesándonos solamente por el resultado final, esto es, la correspondencia entre las posiciones inicial y final de los puntos.

Otra importante transformación del espacio es el desplazamiento paralelo de todos los puntos de una dirección dada a lo largo de una distancia cualquiera. De la figura 6, en la que hemos indicado los puntos Q_1, Q_2, Q_3 , correspondientes a los puntos P_1, P_2, P_3 , se deduce que para hallar el punto correspondiente a otro cualquiera en la

traslación, es suficiente conocer el transformado de un solo punto del espacio.

Hemos definido ya los conceptos de plano, eje y centro de simetría de un cuerpo. A cada uno de estos conceptos corresponde una transformación determinada del espacio: una simetría respecto a un plano, un giro alrededor de una recta y una simetría respecto al centro. Por ejemplo, una simetría respecto a un plano es la transformación que asocia a cada punto del espacio su punto simétrico respecto al plano. Los giros respecto a un eje y las simetrías centrales se definen de modo similar.

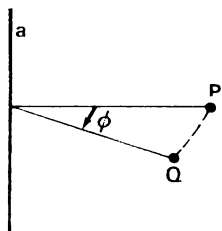


FIG. 5

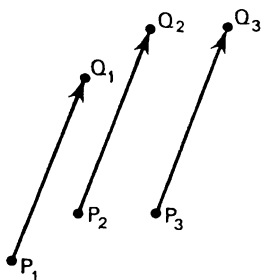


FIG. 6

Hasta aquí hemos hablado de transformaciones del espacio. Las transformaciones correspondientes del plano: giros alrededor de un punto y de un ángulo dado, simetrías respecto a un eje y simetrías centrales, se definen de modo completamente similar (y de forma más intuitiva) en el espacio.

Transformaciones uno a uno

Al discutir las posibles transformaciones de un conjunto en sí mismo debemos en primer lugar observar la diferencia fundamental que existe entre las transformaciones que son uno a uno y las que no lo son.

Una transformación A de un conjunto M se dice que es una transformación uno a uno del conjunto en sí mismo, si además de corresponder a cada elemento de M un único elemento de M (esto es parte de la definición de transformación), para cada elemento y existe un solo elemento x que se transforma en él. Dicho con otras palabras, una transformación es uno a uno si la «ecuación» $xA = y$ tiene una única solución en M cualquiera que sea y en M .

Todas las transformaciones del espacio que hemos considerado aquí —simetrías, giros y traslaciones— son uno a uno, ya que en estos casos no sólo existe para todo punto X un punto en el que éste se transforma, sino que existe también un único punto que se transforma en X .

Es fácil dar ejemplos de la situación contraria; así, la transformación del conjunto de números 1, 2, 3, 4, dada por la tabla

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \end{pmatrix}$$

no es uno a uno, ya que no hay ningún número que se transforme en el 4. La transformación de todos los números naturales 1, 2, 3, $\dots$, dada por la tabla

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & \dots \\ 1 & 1 & 2 & 2 & 3 & 3 & \dots \end{pmatrix}$$

tampoco es uno a uno. Aunque aquí para cada número n existe un número $2n$ que se transforma en él, el número $2n$ no es el único que tiene esta propiedad, ya que el número $2n - 1$ también se transforma en n . Es fácil establecer, para transformaciones dadas por tablas, un criterio que permita saber cuándo son uno a uno. Para ello es necesario y suficiente que la línea inferior de la tabla contenga cada elemento del conjunto una y sólo una vez. En matemáticas se discuten a veces transformaciones que no son uno a uno. Por ejemplo, es bien conocida la importancia de la operación que consiste en proyectar el espacio sobre un plano. Esta transformación no es uno a uno, ya que cada punto de este último es la proyección no de uno, sino de toda una serie de puntos del espacio. Pero en la mayoría de los casos revisten más interés las transformaciones uno a uno; estas transformaciones, en particular, juegan un papel fundamental cuando se consideran procesos físicos en los cuales los elementos del sistema ni se fusionan, ni se destruyen, ni se crean.

En lo sucesivo, al hablar de transformaciones supondremos implícitamente que son uno a uno; también reciben a veces el nombre de permutaciones, en especial cuando se trata de transformaciones de conjuntos finitos.

Para toda transformación (uno a uno) se puede definir una transformación inversa A^{-1} . Si A transforma un elemento arbitrario x de M

en y , entonces la transformación que transforma y en x se llama transformación inversa de A y se denota por A^{-1} . Por ejemplo, si

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix},$$

entonces

$$A^{-1} = \begin{pmatrix} 2 & 3 & 4 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{pmatrix}.$$

Si A es un giro de ángulo ϕ en el espacio alrededor de un eje, entonces A^{-1} es el giro con el mismo eje, de ángulo ϕ en la dirección opuesta, etc.

En ocasiones ocurre que la transformación inversa coincide con la directa. En particular las simetrías respecto a un plano o a un punto tienen esta propiedad. Lo mismo ocurre con la permutación

$$A = \begin{pmatrix} 2 & 1 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}, \text{ ya que } A^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

Observemos que no se puede hablar de la inversa de una transformación si ésta no es uno a uno, ya que un elemento puede ser tal que ningún elemento (o más de uno) se transformen en él.

La definición general de simetría

En matemáticas y sus aplicaciones es muy raro tener que considerar todas las transformaciones de un conjunto dado. De hecho, tampoco es frecuente contemplar los conjuntos como meras colecciones de elementos completamente desconectados los unos de los otros. Y ello es natural, porque los conjuntos que se discuten en matemáticas son imágenes abstractas de colecciones reales, cuyos elementos poseen siempre una infinita variedad de interrelaciones, así como de conexiones con el mundo que está más allá del conjunto en cuestión. Pero en matemáticas es conveniente hacer abstracción de la mayor parte de esas conexiones, preservando y teniendo en cuenta las más esenciales. Esto nos obliga en primer lugar a considerar sólo aquellas transformaciones que conserven las relaciones más importantes entre elementos. Estas transformaciones se denominan a menudo transformaciones admisibles o *automorfismos* con respecto a dichas relaciones entre los elementos del conjunto. Por ejemplo, en el espacio, el concepto de distancia entre dos puntos es muy importante. La presencia de este concepto forja un nexo que consiste

en el hecho de que cada dos puntos guardan una distancia determinada entre ellos. Las transformaciones que conservan estas relaciones son las mismas que conservan invariables las distancias entre cada par de puntos. Estas transformaciones son los llamados «movimientos» del espacio.

Con la ayuda del concepto de automorfismo no es difícil dar una definición general de simetría. Supongamos un cierto conjunto M , en el que se tienen en cuenta ciertas relaciones entre sus elementos, y supongamos que P es una parte de M . Se dice que P es simétrico o invariante respecto a la transformación admisible A de M , si A transforma cada elemento de P en otro elemento de P . Por tanto, una simetría de P se caracteriza por la colección de aquellas transformaciones admisibles del conjunto M que transforman P en sí mismo. El concepto de simetría de un cuerpo en el espacio cae por completo dentro de esta definición. El papel del conjunto M lo desempeña el espacio completo; el de las transformaciones admisibles, los «movimientos», y el papel de P , el cuerpo dado. La simetría de P se caracteriza, por tanto, mediante la colección de movimientos que transforman P en sí mismo.

Las simetrías respecto a un plano, eje o centro, las traslaciones y las rotaciones del espacio alrededor de un eje, que ya hemos discutido, son casos especiales de movimientos, ya que las distancias entre los puntos permanecen evidentemente inalterados bajo dichas transformaciones. Un estudio más detallado de la cuestión muestra que todo movimiento del plano es o una traslación, o un giro alrededor de un centro, o una simetría respecto a un cierto eje, o una combinación de una simetría respecto a un eje con una traslación de dirección paralela a dicho eje. Análogamente, todo movimiento del espacio es o una traslación, o un giro alrededor de un eje o un movimiento helicoidal —es decir, un giro respecto a un eje, combinado con una traslación de dirección paralela a dicho eje— o una simetría respecto a un plano, combinada quizá con una traslación paralela a dicho plano o con un giro respecto a un eje perpendicular a este plano.

Las traslaciones, giros y movimientos helicoidales del espacio reciben el nombre de movimientos propios o movimientos de primera especie. Los demás movimientos (incluidas las simetrías) se conocen como movimientos impropios o movimientos de segunda especie. En el plano, los movimientos de primera especie son las traslaciones y los giros, mientras que las simetrías respecto a una recta y las simetrías combinadas con traslaciones o giros son movimientos de segunda especie.

Es fácil imaginar cómo las transformaciones que son movimientos

de primera especie se pueden obtener a partir de movimientos continuos del plano o del espacio. Los movimientos de segunda especie no se pueden obtener de esta forma, ya que la simetría especular que entra en su formación no le permite.

Se dice a veces que el plano es simétrico en todas sus partes o que todos los puntos del plano son equivalentes. En el lenguaje estricto de las transformaciones, esta afirmación significa que todo punto del plano se puede superponer sobre otro punto cualquiera mediante un «movimiento» apropiado.

Los casos de simetría de cuerpos o figuras discutidos anteriormente están también comprendidos en la definición general de simetría. Por ejemplo, un cuerpo simétrico respecto a un plano α coincide consigo mismo después de efectuar una simetría respecto a dicho plano; un cuerpo que tiene un centro de simetría O se transforma en sí mismo mediante una simetría de centro O . Por tanto, el grado de simetría de un cuerpo o de una figura espacial se puede caracterizar completamente mediante la colección de todos los movimientos espaciales de primera o segunda especie que lo transforman en sí mismo. Cuanto mayor y más diversa es la colección de movimientos, mayor es el grado de simetría del cuerpo o figura. Si, en particular, esta colección no contiene ningún movimiento, salvo la transformación idéntica, se dice entonces que el cuerpo es asimétrico.

El grado de simetría de un cuadrado en el plano se caracteriza por la colección de movimientos del plano que transforman el cuadrado en sí mismo. Pero, si el cuadrado coincide consigo mismo, lo mismo debe ocurrir con el punto de intersección de las diagonales. Por tanto, los movimientos buscados dejan invariante el centro del cuadrado y son, por consiguiente, giros alrededor del centro o simetrías

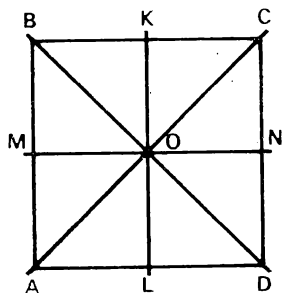


FIG. 7

respecto a rectas que pasan por dicho punto. De la figura 7 se puede fácilmente deducir que el cuadrado $ABCD$ es simétrico respecto a los giros de centro O y ángulo un múltiplo de 90° , así como respecto a las simetrías de ejes las diagonales AC , BD y las rectas KL , MN . Estos ocho movimientos caracterizan la simetría del cuadrado.

La colección de las simetrías de un rectángulo se reduce a un giro de 180° alrededor del centro del rectángulo y a las simetrías de ejes

las rectas que unen los puntos medios de los lados; el conjunto de las simetrías de un paralelogramo (Fig. 8) consta solamente de los giros alrededor del centro de ángulos múltiplos de 180° , esto es, de una simetría central y de la transformación identidad.

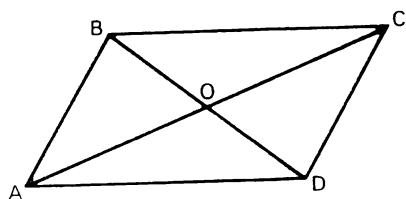


FIG. 8

Hemos dado anteriormente un ejemplo algebraico de simetría; dijimos que el concepto de simetría de un polinomio de varias variables tenía también sentido.

Discutamos ahora cómo se puede caracterizar la simetría de un polinomio.

Diremos que se ha efectuado una permutación de variables

$$A = \begin{pmatrix} x_1, x_2, \dots, x_n \\ x_{i_1}, x_{i_2}, \dots, x_{i_n} \end{pmatrix}$$

o más brevemente

$$A = \begin{pmatrix} 1, 2, \dots, n \\ i_1, i_2, \dots, i_n \end{pmatrix}$$

en el polinomio $F(x_1, x_2, \dots, x_n)$, si en todas partes donde aparece x_1 se ha sustituido por x_{i_1} , x_2 por x_{i_2} , etc. El polinomio obtenido así lo denotaremos por FA , si $F = x_1^2 - 2x_2 + x_3 - x_4$,

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}, \text{ entonces } FA = x_3^2 - 2x_1 + x_4 - x_2.$$

La simetría del polinomio dado se caracteriza por la colección de permutaciones de variables que al ser efectuada sobre el polinomio lo dejan invariable. Por ejemplo, la simetría del polinomio $x_1^3 + 2x_2 + x_3^3 + 2x_4$ se caracteriza por las permutaciones

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix},$$

y la simetría del polinomio $x_1^3 + 2x_2 + x_3^3 + x_4$ se caracteriza por las dos permutaciones:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix} \quad \text{y} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix}$$

§3. Grupos de transformaciones

Producto de transformaciones

Al estudiar las propiedades de las transformaciones es fácil observar que algunas de ellas se pueden obtener a partir de otras. Por ejemplo, el movimiento helicoidal se compone de un giro alrededor de un eje y una traslación de dirección paralela a dicho eje. Este proceso de formación de transformaciones a partir de otras dadas se llama producto de transformaciones. Si aplicamos a un elemento x arbitrario de M una transformación A y a continuación aplicamos la transformación B al elemento xA , obtenemos el elemento $(xA)B$. La transformación que convierte a x en $(xA)B$ se llama producto de A y B y se denota por AB . Así, pues, por definición, tenemos

$$x(AB) = (xA)B.$$

Ejemplo:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{pmatrix}.$$

Puesto que la primera permutación transforma el 1 en el 2 y la segunda, el 2 en el 4, la permutación resultante debe llevar el 1 sobre el 4, y así sucesivamente. He aquí algunos ejemplos más:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix};$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 4 & 2 \end{pmatrix}; \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 1 & 3 \end{pmatrix}.$$

Los dos últimos ejemplos muestran que el producto de transformaciones es, por así decirlo, una operación no conmutativa: su resultado depende del orden de los factores. Esto se verifica también fácilmente para movimientos del plano. Supongamos que A es un giro de centro el origen y amplitud 90° , y B una traslación de longitud unidad paralela al eje x y de sentido positivo.

Halleemos la imagen de O bajo las transformaciones AB y BA . Tenemos, por definición (Fig. 9):

$$O(AB) = (OA)B = OB = M,$$

$$O(BA) = (OB)A = MA = N,$$

es decir, $AB \neq BA$.

Para comprender mejor la naturaleza geométrica de la transformación BA , consideremos el punto P . Tenemos

$$P(BA) = (PB)A = QA = P,$$

esto es, la transformación BA deja invariante el punto P . Partiendo de esto, es fácil probar que BA no es más que un giro de amplitud 90° alrededor de P . Análogamente,

$$Q(AB) = (QA)B = PB = Q,$$

y AB es una rotación de centro Q y amplitud 90° .

El producto de movimientos del plano o del espacio sigue en general reglas más bien complicadas. Sin embargo, en dos importantes casos son muy sencillas.

Primero, si multiplicamos giros del plano alrededor del mismo punto o giros en el espacio alrededor de una misma recta, con ángulo ϕ y ψ , la transformación producto es el correspondiente giro de ángulo $\phi + \psi$. En segundo lugar, si multiplicamos traslaciones caracterizadas por los vectores $\vec{MN}$ y $\vec{NP}$, el producto es también una traslación de vector $\vec{MP}$, es decir, la suma de los vectores originales.

El mismo término «producto» de transformaciones indica una cierta analogía con el producto de números. Sin embargo, esta analogía no es completa. Por ejemplo, el producto de números posee la propiedad conmutativa, y sin embargo el producto de transformaciones, como acabamos de ver, no la tiene. La segunda ley fundamental de la aritmética, la ley asociativa, se cumple también en el caso de las transformaciones. De hecho, para tres transformaciones arbitrarias de un conjunto M tenemos la ecuación $A(BC) = (AB)C$.

Para probarlo basta tomar un elemento m de M , entonces

$$m[A(BC)] = (mA)(BC) = [(mA)B]C = [m(AB)]C = m[(AB)C].$$

La ley asociativa permite hablar solamente del producto $A(BC) = (AB)C = ABC$ sin necesidad de distinguir entre los dos productos $A(BC)$ y $(AB)C$ de las transformaciones A, B, C . La misma ley

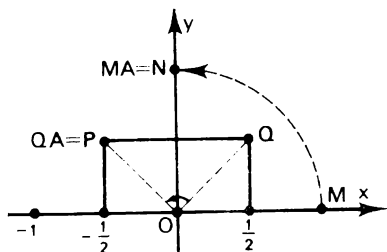


FIG. 9

muestra que el producto de cuatro o más transformaciones no depende de la distribución de los paréntesis.

Además, entre las transformaciones existe una, que juega el papel del número 1; es la transformación identidad E , que deja invariable todo elemento de M . Es obvio que $AE = EA = A$, cualquiera que sea la transformación A .

Mencionemos el hecho siguiente: el producto de transformaciones uno a uno es también uno a uno. Pues para hallar el elemento x de M que se transforma en un elemento dado a bajo AB , es suficiente hallar el elemento x_1 que B transforma en a , y a continuación el elemento x_2 que se transforma mediante A en x_1 . Puesto que $x_2(AB) = (x_2A)B = x_1A = a$, entonces x_2 es el elemento x que buscábamos.

El producto de una transformación A por su inversa A^{-1} es la transformación identidad; es decir,

$$AA^{-1} = A^{-1}A = E.$$

Esto se sigue inmediatamente de la definición de transformación inversa.

El ejemplo, discutido anteriormente, de la multiplicación de una traslación por un giro en el plano muestra que las propiedades de un producto no siempre se deducen fácilmente de las de los factores. Sin embargo, el producto de transformaciones de la forma $C = B^{-1}AB$ es una excepción importante. Las propiedades de C se relacionan de una manera muy simple con las de A y B . Sea m un elemento de M transformado en el elemento n por A ; entonces el elemento mB , transformado de m por B , es llevado por C sobre nB , imagen de n por B .

$$\text{Demostración: } (mB)B^{-1}AB = mAB = nB.$$

Se dice que la transformación $B^{-1}AB$ se obtiene de A mediante una transformación por B , o que es la conjugada de A por B .

Transformemos, por ejemplo, un giro P_0 de centro O por medio de una traslación V (en el plano). Según la regla precedente, para hallar las parejas de posiciones inicial y final de los puntos para el movimiento transformado $C = V^{-1}P_0V$, debemos transformar por V las parejas correspondientes a la transformación P_0 . Puesto que el punto O en el giro P_0 permanece invariable (Fig. 10), el punto OV

también permanecerá invariable bajo C . Además, si P_0 lleva un punto M sobre N , la transformación C transformará MV en NV . De la figura 10 se deduce que C es un giro de centro el punto OV y del mismo ángulo ϕ que el giro P_0 .

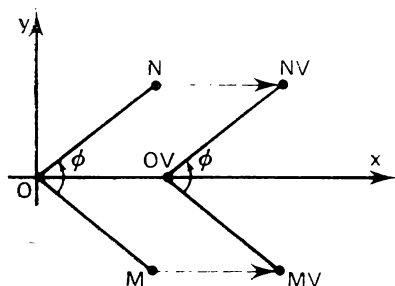


FIG. 10

Del mismo modo se puede probar que si transformamos una traslación caracterizada por el vector $\overrightarrow{MN}$ por medio de un giro P_0 de ángulo ϕ , obtenemos de nuevo una traslación caracterizada por un vector diferente.

La regla anterior para hallar la transformación $B^{-1}AB$ se puede formular de una manera muy elegante cuando las transformaciones vienen dadas por tablas. Supongamos que

$$A = \begin{pmatrix} 1 & 2 & \cdots & n \\ a_1 a_2 & \cdots & a_n \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 & \cdots & n \\ b_1 b_2 & \cdots & b_n \end{pmatrix},$$

entonces

$$B^{-1}AB = \begin{pmatrix} b_1 b_2 & \cdots & b_n \\ 1 & 2 & \cdots & n \end{pmatrix} \begin{pmatrix} 1 & 2 & \cdots & n \\ a_1 a_2 & \cdots & a_n \end{pmatrix} \begin{pmatrix} 1 & 2 & \cdots & n \\ b_1 b_2 & \cdots & b_n \end{pmatrix} = \begin{pmatrix} b_1 & b_2 & \cdots & b_n \\ b_{a_1} b_{a_2} & \cdots & b_{a_n} \end{pmatrix};$$

es decir, para transformar una permutación A mediante otra permutación B debemos someter todos los elementos de las filas superior e inferior de A a la transformación indicada por B . Por ejemplo, si

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 4 & 1 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 1 & 3 & 4 \end{pmatrix},$$

entonces

$$B^{-1}AB = \begin{pmatrix} 1B & 2B & 3B & 4B & 5B \\ 3B & 5B & 4B & 1B & 2B \end{pmatrix} = \begin{pmatrix} 2 & 5 & 1 & 3 & 4 \\ 1 & 4 & 3 & 2 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 1 & 2 & 5 & 4 \end{pmatrix},$$

Observamos que, aunque en general el producto de dos transformaciones depende del orden de los factores, en algunos casos los productos AB y BA pueden coincidir. Se dice entonces que las transformaciones A y B son permutables o que conmutan. Si $AB = BA$, entonces

$$B^{-1}AB = B^{-1}BA = A.$$

Así, pues, la transformación de una permutación por medio de otra que conmuta con ella no altera la permutación dada.

Grupos de transformaciones

El conjunto de las transformaciones que caracterizan la simetría de una figura no puede ser arbitrario, sino que debe necesariamente cumplir las siguientes propiedades:

1. El producto de dos transformaciones del conjunto pertenece también al conjunto.
2. La transformación identidad pertenece al conjunto.
3. Si una transformación pertenece al conjunto, su inversa también pertenece al conjunto.

Estas propiedades son muy importantes en el estudio de las transformaciones; a consecuencia de ello, todo conjunto de transformaciones uno a uno que tenga estas tres propiedades recibe el nombre de *grupo de transformaciones*, caracterice o no la simetría de alguna figura.

Desde el punto de vista del álgebra, las propiedades 1-3 son muy importantes, ya que permiten formar, partiendo de ciertas transformaciones $A, B, C, \dots$ del conjunto dado, otras nuevas de la forma $ABAC, A^{-1}BCB^{-1}$, etc., garantizándonos las propiedades 1-3 que todas las transformaciones así obtenidas caen dentro de los límites del conjunto de transformaciones dado.

El número de transformaciones que forman un grupo se llama *orden del grupo*; puede ser finito o infinito.

De acuerdo con esto, los grupos se dividen en finitos e infinitos. Hemos discutido anteriormente el grupo de simetría de un cuadrado en el plano. Este grupo consistía en un conjunto de ocho transformaciones. Por otra parte,

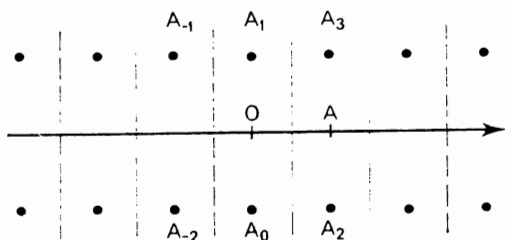


FIG. 11

el conjunto de los infinitos puntos A_i ilustrados en la figura 11 se transforma en sí mismo mediante los siguientes movimientos del plano: traslaciones a lo largo del eje OA en una u otra dirección de longitud

igual a un múltiplo de OA ; simetrías de ejes las líneas de puntos; simetría de eje OA . Por tanto, es evidente que el grupo de simetrías de esta figura es infinito.

La colección de las transformaciones que conservan un cierto objeto, es decir, que caracterizan su simetría, es siempre un grupo. Este método de dar grupos en forma de simetría es de la máxima importancia. Mediante este principio se pueden obtener grupos muy importantes. Los principales son los grupos de movimientos del plano y del espacio. Los grupos de simetría de los poliedros regulares tienen también un gran interés. Se sabe que en el espacio hay en total cinco tipos de poliedros regulares (con 4, 6, 8, 12 y 20 caras). Si tomamos un poliedro regular arbitrario y consideramos todos los movimientos del espacio que llevan al poliedro a coincidir consigo mismo, obtenemos un grupo, a saber el grupo de simetría del poliedro. Si en lugar de todos los movimientos consideramos sólo aquellos de primera especie que hacen coincidir el poliedro consigo mismo, obtenemos un grupo que es parte del grupo total de las simetrías del poliedro. Este grupo se llama grupo de rotaciones del poliedro. Puesto que en una superposición del poliedro consigo mismo su centro permanece invariable, todos los movimientos que entran en el grupo de simetrías del poliedro dejan el centro invariante y no pueden ser, por tanto, más que giros cuyos ejes pasen por el centro, o simetrías respecto a planos que pasen por el centro, o finalmente simetrías respecto a dichos planos, combinadas con giros cuyos ejes sean perpendiculares a estos planos y pasen por el centro.

Con la ayuda de estas notas es fácil encontrar todos los grupos de simetría y de rotaciones de los poliedros regulares. En la tabla 1 se dan los órdenes de los grupos de simetrías y grupos de rotaciones de los poliedros regulares. Todos estos grupos son finitos.

TABLA 1

Número de caras	4	6	8	12	20
Orden del grupo de simetría	24	48	48	120	120
Orden del grupo de rotaciones	12	24	24	60	60

Grupos de permutaciones

De todos los grupos de transformaciones, los primeros en ser estudiados, históricamente, en matemáticas fueron los grupos de permutaciones de polinomios en las variables $x_1, x_2, \dots, x_n$. El estudio de estos grupos esta íntimamente relacionado con el problema de resolver ecuaciones de grado superior mediante radicales. Evidentemente, la colección de todas aquellas permutaciones de las variables que no alteran el valor de uno o varios polinomios forma un grupo. Los polinomios que permanecen invariables bajo cualquier permutación se llaman polinomios simétricos. Por ejemplo, $x_1 + x_2 + \dots + x_n$ es un polinomio simétrico. De acuerdo con esto, la colección de todas las permutaciones de un conjunto dado de variables recibe el nombre de grupo *simétrico* de las permutaciones de dicho conjunto.

El número de variables que se permutan se llama grado del grupo simétrico. En lugar de las permutaciones de las variables $x_1, \dots, x_n$, podemos considerar solamente las permutaciones de los números $1, 2, \dots, n$. Puesto que toda permutación de estos números se puede escribir de la forma

$$\begin{pmatrix} 1 & 2 & \dots & n \\ a_1 & a_2 & \dots & a_n \end{pmatrix},$$

donde $a_1, a_2, \dots, a_n$ son los números $1, 2, \dots, n$ escritos en un cierto orden, el número de todas las permutaciones de n elementos, esto es, el orden del grupo simétrico, es igual a $n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot n$. Este orden aumenta muy rápidamente con n ; el grupo de permutaciones de 10 variables es ya de orden 3.628.800.

Consideremos el polinomio

$$\begin{aligned} F(x_1, x_2, \dots, x_n) = \\ = (x_2 - x_1)(x_3 - x_1) \cdots (x_n - x_1)(x_3 - x_2) \cdots (x_n - x_2) \cdots (x_n - x_{n-1}) (1) \end{aligned}$$

Es evidente que toda permutación de las variables, o bien deja invariable el valor del polinomio F , o bien cambia solamente su signo. Las permutaciones de la primera clase se llaman pares. Las que cambian el signo de F se llaman impares. El conjunto de las permutaciones pares que forman el grupo de simetrías del polinomio (1) recibe el nombre de grupo alternado.

El producto de dos permutaciones pares es par, ya que las permutaciones pares forman grupo. El producto de dos permutaciones impares es una permutación par. Pues si A y B son permutaciones impares, entonces:

$$FAB = (FA)B = (-F)B = -(-F) = F.$$

De la misma forma se puede probar que el producto de una permutación par y otra impar es una permutación impar y que la permutación inversa de una permutación par o impar es de la misma paridad.

Un ejemplo de permutación impar es

$$S = \begin{pmatrix} 1, 2, 3, \dots, n \\ 2, 1, 3, \dots, n \end{pmatrix},$$

que intercambia los elementos 1 y 2.

Descomposición de una permutación en ciclos

En el estudio de los grupos de permutaciones es muy conveniente representar las permutaciones en forma de productos de ciclos. Por definición, el símbolo $(m_1, m_2, \dots, m_k)$ denota la permutación que lleva m_1 sobre m_2 , m_2 sobre m_3 , $\dots$, m_{k-1} sobre m_k , y m_k de nuevo sobre m_1 . Por ejemplo, si consideramos permutaciones de los números 1, 2, 3, 4, 5, entonces

$$(1, 2, 3, 4, 5) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 1 \end{pmatrix}, \quad (3, 5) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 5 & 4 & 3 \end{pmatrix}.$$

Una permutación de la forma $(m_1, m_2, \dots, m_k)$ se llama *cíclica* o *ciclo de longitud k*, y $m_1, m_2, \dots, m_k$ se dice que son los elementos del ciclo. La permutación unidad se puede escribir en forma de ciclos $(1) = (2) = \dots$ de longitud 1. Los ciclos de longitud 2 se llaman *transposiciones*. Si permutamos los elementos de un ciclo según un orden cíclico, obtenemos la misma permutación; por ejemplo $(1, 2, 3) = (2, 3, 1) = (3, 2, 1)$; $(5, 6) = (6, 5)$.

Es fácil comprobar que los ciclos sin elementos comunes, por ejemplo, $(2, 3)$ y $(1, 4, 5)$, son permutables, de modo que al multiplicarlos no es necesario tener en cuenta el orden en que entran los factores.

La importancia de los ciclos en la teoría general se basa en el siguiente teorema: toda permutación se puede representar en forma

de un producto de ciclos sin elementos comunes, y esta representación es única salvo el orden de los factores.

La demostración del teorema es inmediata a partir del método de esta representación. Supongamos que se quiere descomponer la permutación

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 3 & 2 & 1 \end{pmatrix}.$$

Vemos que A transforma el 1 en el 4, el 4 en el 3, el 3 en el 6, y el 6 en el 1. Tenemos, por tanto, como primer factor el $(1, 4, 3, 6)$. Consideremos ahora el 2; observemos que A lo transforma en el 5, y éste se transforma en el 2 de nuevo. Por tanto, el segundo factor es $(2, 5)$. Puesto que hemos considerado ya la transformación de todos los números, tenemos que

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 3 & 2 & 1 \end{pmatrix} = (1, 4, 3, 6)(2, 5). \quad (2)$$

También es posible descomponer las permutaciones en ciclos con elementos comunes, pero en este caso la descomposición no es única. Por ejemplo,

$$\begin{aligned} (a_1, a_2, \dots, a_n) &= (a_1, a_2)(a_1, a_3) \cdots (a_1, a_n) = \\ &= (a_2, a_3)(a_2, a_4) \cdots (a_2, a_n)(a_2, a_1). \end{aligned} \quad (3)$$

Probemos que todo ciclo de longitud 2 es una permutación impar. Lo hemos visto ya para el ciclo $(1, 2)$. Pero todo ciclo (i, j) es $S^{-1}(1, 2)S$, siendo S una permutación arbitraria

$$\begin{pmatrix} 1, 2, \dots \\ i, j, \dots \end{pmatrix}$$

que lleva el 1 sobre el i y el 2 sobre el j . La permutación $S^{-1}(1, 2)S$ es impar, ya que $(1, 2)$ es impar, y S^{-1} y S son simultáneamente pares o impares.

De acuerdo con (3), un ciclo de longitud $m + 1$ se puede escribir como producto de m permutaciones impares. Por tanto, un ciclo de longitud $m + 1$ es una permutación par, si $m + 1$ es par, e impar, si $m + 1$ es impar. Esto permite calcular rápidamente la paridad de

las permutaciones una vez conocida su descomposición en ciclos. Concretamente, la permutación

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 3 & 2 & 1 \end{pmatrix}$$

es par, ya que por (2) se ve que es el producto de dos permutaciones pares.

Subgrupos

Una parte de un grupo que a su vez es un grupo se dice que es un *subgrupo* del grupo dado. Así, el grupo alternado de permutaciones de las variables $x_1, x_2, \dots, x_n$ es un subgrupo del grupo simétrico. El conjunto de todos los movimientos propios de un plano es un subgrupo del grupo de todos los movimientos propios e improprios del plano.

Desde un punto de vista formal, la transformación identidad forma por sí sola un subgrupo. Todo grupo se puede considerar, igualmente, un subgrupo de sí mismo. Pero casi siempre los grupos contienen muchos otros subgrupos además de estos que podríamos llamar triviales. El conocimiento de todos los subgrupos de un grupo da una idea bastante completa de la estructura interna de éste.

Uno de los métodos más corrientes de formar subgrupos es el de especificar lo que se denomina generadores del subgrupo.

Sean $A_1, A_2, \dots, A_m$ transformaciones arbitrarias pertenecientes a un grupo G . El conjunto H de todas las transformaciones que se pueden obtener multiplicando las transformaciones dadas y sus inversas es un subgrupo, pues la transformación unidad pertenece al conjunto, ya que se puede escribir en la forma $A_1^{-1}A_1$. Además, si las transformaciones B y C son productos del tipo mencionado, multiplicando estos productos obtenemos una representación para BC también en forma de producto. Finalmente, si B se expresa en forma de producto, por ejemplo $B = A_1^{-1}A_2A_1A_2^{-1}$, entonces B^{-1} también se puede escribir en forma de producto, ya que $B^{-1} = A_2A_1^{-1}A_1^{-1}A_2^{-1}A_1$.

Es evidente que H es un subgrupo de G y recibe el nombre de subgrupo generado por las transformaciones $A_1, \dots, A_m$; estas transformaciones se llaman los generadores de H . Puede ocurrir que H coincida con G , y en este caso se dice que $A_1, \dots, A_m$ son los generadores de todo el grupo G . Es fácil verificar por medio de ejemplos que un mismo subgrupo puede ser engendrado por distintos sistemas de generadores.

Un subgrupo generado por una sola transformación A se llama cíclico. Sus elementos son

$$E, A, AA, AAA, \dots, A^{-1}, A^{-1}A^{-1}, A^{-1}A^{-1}A^{-1}, \dots,$$

que de modo natural reciben el nombre de potencias de A . Se escriben así:

$$E = A^0, A = A^1, AA = A^2, \dots, A^{-1}A^{-1} = A^{-2}, \\ A^{-1}A^{-1}A^{-1} = A^{-3}, \dots$$

Es fácil probar, igual que en aritmética, que

$$A^m A^n = A^{m+n} \quad \text{y} \quad (A^m)^n = A^{mn}. \quad (4)$$

Una transformación se llama *periódica* si alguna de sus potencias es la transformación identidad. El número positivo más pequeño al que hay elevar la transformación periódica para obtener la identidad recibe el nombre de *orden de la transformación*. Se dice también que una transformación que no es periódica es de orden infinito.

Consideremos algunos ejemplos. Sea A un giro (en el plano) de centro O y ángulo $360^\circ/n$, donde n es un entero positivo mayor que 1. Entonces A^2 es un giro de centro O y ángulo $2(360^\circ/n)$, A^3 de $3(360^\circ/n)$, y A^{n-1} un giro de $(n-1)(360^\circ/n)$; por último, A^n es un giro de 360° , es decir, la transformación identidad. Esto prueba que un giro de ángulo $360^\circ/n$ es una transformación periódica de orden n .

Sea A una traslación del plano. Entonces $A^2, A^3, \dots$ son también traslaciones de la misma dirección y sentido, pero de longitud doble, triple, etc. Por tanto, ninguna potencia positiva de A puede ser la identidad, y en consecuencia el orden de A es infinito.

Los elementos del grupo cíclico generado por A son

$$\dots, A^{-2}, A^{-1}, E, A, A^2, \dots \quad (5)$$

Si A es una transformación de orden infinito, todas las transformaciones de la sucesión (5) son distintas y el grupo es infinito. Pues de otra forma tendríamos una ecuación de la forma $A^k = A^l$ ($k < l$), de donde $A^{l-k} = E$ ($l-k > 0$), lo cual contradice el hecho de que A no es periódica.

Supongamos ahora que A es una transformación periódica de orden m . Entonces

$$A^m = E, A^{m+1} = A, A^{m+2} = A^2, \dots, A^{m-1} = A^{-1}, A^{m-2} = A^{-2}, \dots;$$

es decir, la sucesión (5) consta precisamente de las transformaciones $E, A, A^2, \dots, A^{m-1}$ repetidas periódicamente. Son distintas entre sí, ya que si $A^k = A^l$ ($0 \leq k < l < m$), entonces $A^{l-k} = E$ ($0 < l-k < m$), en contradicción con la elección de m . En consecuencia, el grupo cíclico engendrado por una transformación de orden m contiene exactamente m transformaciones distintas.

Un grupo en el que todos los elementos conmutan dos a dos se llama grupo conmutativo o *abeliano*, en honor al matemático Abel, que descubrió la gran importancia de estos grupos en la teoría de ecuaciones.

Las fórmulas (4) muestran que las potencias de una transformación conmutan siempre unas con otras: $A^m A^n = A^n A^m = A^{m+n}$. Por tanto, los grupos cíclicos son siempre abelianos.

En la teoría de grupos, como la multiplicación no tiene por qué ser conmutativa, hay que hablar de dos divisiones: por la derecha y por la izquierda. La solución de la ecuación $Ax = B$, donde A y B son transformaciones dadas, se llama cociente por la derecha, y la solución de la ecuación $yA = B$, cociente por la izquierda de B por A . Si multiplicamos ambos miembros de la primera ecuación por A^{-1} a la izquierda y ambos miembros de la segunda por A^{-1} a la derecha, obtenemos: $x = A^{-1}B$, $y = BA^{-1}$. Así, pues, podemos considerar las transformaciones $A^{-1}B$ o BA^{-1} como el «cociente» entre B y A .

En numerosos ejemplos hemos visto que, en general, $AB \neq BA$. El «cociente» $(AB)(BA)^{-1}$ o $(BA)^{-1}(AB)$ se puede tomar como una medida de la no conmutatividad de las permutaciones A y B . La segunda de estas expresiones, $(BA)^{-1}(AB) = A^{-1}B^{-1}AB$, se llama *conmutador* de A y B y se denota por (A, B) . De la fórmula

$$(A, B) = A^{-1}B^{-1}AB$$

se sigue que el conmutador se puede considerar como el cociente de «dividir» la transformación conjugada $B^{-1}AB$ por A .

Por ejemplo, si A es una traslación del plano, una transformación conjugada será también una traslación, y el cociente de dos traslaciones también lo es. Por tanto, el conmutador de una traslación y de un movimiento arbitrario en el plano es una traslación. Sea ahora A un giro de centro O y ángulo ϕ , y B una traslación o un giro. Entonces

la transformación conjugada es otro giro de ángulo ϕ pero de centro O' . Por tanto, en este caso, el conmutador (A, B) es el producto de un giro de centro O y ángulo $-\phi$ por otro giro de centro O' y ángulo ϕ . De la figura 12 se deduce que la transformación resultante es una traslación de módulo $2 \cdot \overline{OO'}$ sen $\phi/2$ y cuya dirección forma un ángulo $\pi/2 - \phi/2$ con el segmento $\overline{OO'}$.

De este modo hemos llegado a la interesante conclusión de que en el plano, el conmutador de cada dos movimientos de primera especie es una traslación o la transformación identidad. Puesto que $(A, B) = E$ significa que $AB = BA$, todo grupo no conmutativo de movimientos de primera especie en el plano contiene traslaciones.

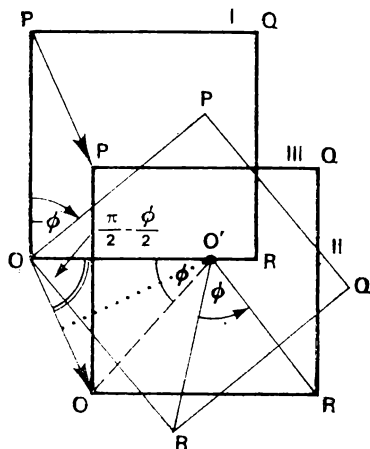


FIG. 12

El subgrupo engendrado por los conmutadores de todos los elementos de un grupo G recibe el nombre de *subgrupo conmutador* o *grupo derivado* de G . Recordando definiciones anteriores, podemos decir que el grupo derivado de G consta precisamente de todos los elementos que se pueden escribir como productos de conmutadores. Puesto que en el plano el conmutador de cualesquiera dos movimientos de primera especie es una traslación, y los productos de traslaciones son también traslaciones, podemos decir que el grupo derivado del grupo de los movimientos de primera especie del plano consta solamente de traslaciones.

El grupo derivado de un grupo abeliano consta solamente de la transformación identidad, ya que como $AB = BA$ tenemos $(A, B) = E$.

Sea G el grupo simétrico de todas las permutaciones de los números $1, 2, \dots, n$. Probemos que el conmutador de cualesquiera dos permutaciones es siempre una permutación par. Efectivamente, las permutaciones AB, BA y en consecuencia $(BA)^{-1}$ tienen siempre la misma paridad; entonces el conmutador $(A, B) = (BA)^{-1}(AB)$, como producto de permutaciones de la misma paridad, es una permutación par.

Hemos visto que el grupo derivado del grupo simétrico consta solamente de permutaciones pares. Es fácil demostrar que, de hecho, coincide con todo el grupo alternado.

El grupo derivado de un grupo G se denota frecuentemente por G' ; el grupo derivado del grupo derivado de G se llama segundo derivado y se denota por G'' . Repitiendo este proceso se puede definir el grupo derivado de cualquier orden de G .

Si entre los grupos derivados de un grupo G al menos uno (y por tanto todos los siguientes) consta solamente de la transformación unidad, se dice que el grupo G es *resoluble*. Este nombre ha nacido en la teoría de ecuaciones, donde la *resolubilidad* de un grupo corresponde a la resolubilidad de una ecuación mediante radicales. El grupo de movimientos de primera especie del plano es soluble, ya que su segundo grupo derivado es la identidad. Los grupos simétricos de grados 2, 3 y 4 son resolubles porque sus grupos derivados primero, segundo y tercero, respectivamente, son la identidad. Por el contrario, los grupos de grado 5 y superior a 5 no son resolubles, ya que se puede demostrar que su segundo grupo derivado coincide con el primero y es diferente de la identidad.

§4. Grupos de Fedorov (Grupos cristalográficos)

Los grupos de simetría de figuras planas finitas

Como ya hemos visto, la simetría de una figura o un cuerpo se caracteriza por el grupo de movimientos del plano que transforman la figura en sí misma.

Los grupos de simetría de figuras planas finitas son los más fáciles de hallar ¹. Consideremos una figura finita y un movimiento A que la transforma en sí misma. Entonces el centro de gravedad O de la figura debe permanecer invariable en esa transformación; esto es, A debe ser o bien un giro de centro O , o bien una simetría respecto a un eje que pasa por O . Así, pues, el grupo de simetría de una figura plana finita sólo puede constar de giros alrededor del centro de gravedad de la figura y de simetrías respecto a ejes que pasan por dicho centro.

Discutamos algunos de los diferentes casos que pueden surgir en el estudio del grupo de simetrías de figuras planas finitas.

¹ La finitud se debe entender en el sentido de que la figura completa está en una parte acotada del plano, por ejemplo, dentro de un círculo.

1. El grupo de simetría K_1 consta solamente de la transformación unidad (identidad). Este es el grupo de simetría de cualquier figura asimétrica (Fig. 13).

2. El grupo de simetría K_2 consta de la transformación unidad y de una simetría respecto a un único eje (Fig. 14).

Observemos que, si un grupo K contiene dos simetrías respecto a dos ejes, a y b , que pasan por O y forman un ángulo ϕ entre sí, el producto de estas dos simetrías es un giro de centro O y ángulo 2ϕ (Fig. 15). Por tanto, está claro que el grupo K_2 es el único grupo de simetría que no contiene giros.

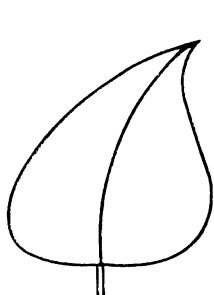


FIG. 13

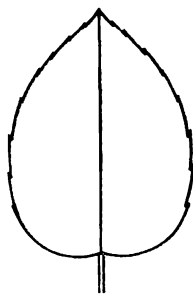


FIG. 14

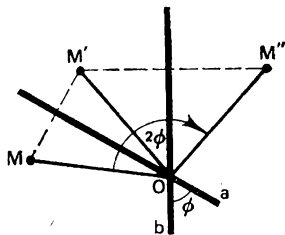


FIG. 15

3. El grupo de simetría K_3 contiene solamente giros, pero no de ángulos arbitrariamente pequeños. En ese caso existe, entre todos los giros, uno de ángulo de giro positivo más pequeño que los demás. Sea este ángulo α° . Probemos que el ángulo de cualquier otro giro del grupo es un múltiplo de α° . Denotamos por β° el ángulo de otro giro y tratamos de hallar el número entero h para el que $h\alpha^\circ \leq \beta^\circ < (h+1)\alpha^\circ$, o lo que es lo mismo, $0 \leq \beta^\circ - h\alpha^\circ < \alpha^\circ$. El grupo K_3 , que contiene giros de ángulos α° y β° , debe contener también un giro de ángulo $\beta^\circ - h\alpha^\circ$. Pero $0 \leq \beta^\circ - h\alpha^\circ < \alpha^\circ$, y el grupo no contiene giros positivos de ángulo menor que α° . Por tanto, $\beta^\circ - h\alpha^\circ = 0$; es decir, $\beta^\circ = h\alpha^\circ$. En particular, como el grupo K_3 contiene el giro de ángulo 360° , tenemos $n\alpha^\circ = 360^\circ$, para un cierto entero n , y de aquí $\alpha^\circ = 360/n$.

Así, pues, el grupo K_3 consta de los giros de ángulos $0^\circ, 360^\circ/n,$

$2(360^\circ/n), \dots, (n-1)(360^\circ/n)$. Dando a n los valores $2, 3, 4, \dots$, obtenemos todos los tipos de grupos K_3 . En la figura 16 hemos ilustrado figuras cuyos grupos de simetría constan solamente de giros de centro O y ángulos múltiplos de $360^\circ/n$, para $n = 19, n = 3$.

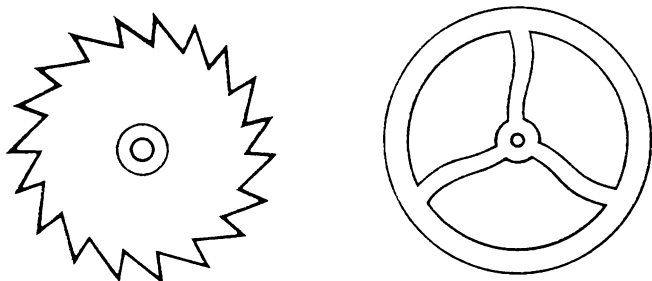


FIG. 16

4. El grupo de simetría K_4 consta sólo de giros, que pueden ser arbitrariamente pequeños. Todo giro de ángulo arbitrario α se puede construir con cualquier grado de aproximación mediante giros de K_4 . Desde luego, aquí solamente nos interesan las figuras cerradas, esto es, figuras tales que incluyen sus puntos fronteras (ver capítulo 16, §9).

Es fácil establecer que para figuras cerradas K_4 contiene giros de cualquier ángulo ϕ . Este es el caso de las figuras de simetría circular dirigida (ilustrado mediante una circunferencia, una corona circular, etcétera) provistas de un sentido concreto de dirección (Fig. 17). Aquí no sólo es la figura

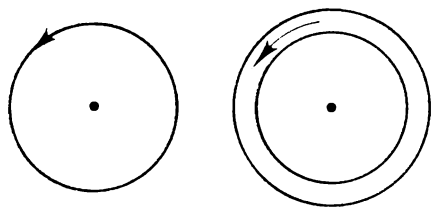


FIG. 17

lo que debe coincidir bajo una transformación admisible, sino también el sentido de dirección, y esto excluye las simetrías respecto a una recta.

Quedan ahora por discutir los casos mixtos en que el grupo de simetría K contiene giros y simetrías respecto a un eje. Sin entrar en demostraciones, que, por otro lado, serían muy simples, enuncia-

mos solamente el resultado: además de los grupos K_1 a K_4 existen solamente los dos siguientes tipos de grupos.

5. El grupo de simetría K_5 consta de n simetrías respecto a ejes que pasan por O y dividen el plano en $2n$ partes iguales, y de giros cuyos ángulos son múltiplos de $360^\circ/n$. Por ejemplo, los polígonos regulares de n lados (Fig. 18) tienen tales grupos de simetría.

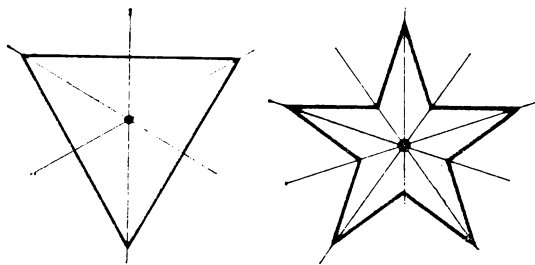


FIG. 18

6. El grupo de simetría K_6 consta de todos los giros de centro O y de todas las simetrías respecto a ejes que pasan por el centro O . Este es el caso de la simetría circular completa, que se puede ilustrar mediante la simetría de una circunferencia o de una corona circular no orientadas.

Grupos de simetría de figuras planas infinitas

La tarea de hallar todos los posibles grupos de simetría de figuras planas infinitas es más complicada. Por supuesto, en la práctica jamás se da el plano infinito completo. Sin embargo, a menudo sucede que un trozo de plano está cubierto de figuras tan pequeñas que, comparado con ellas, se puede considerar infinitamente grande. Por ejemplo, la superficie plana de una pieza de acero finamente pulida aparece cubierta de figuras de dimensiones microscópicas. La regularidad de estas figuras indica la homogeneidad de la estructura interna del metal.

Otros posibles ejemplos son los dibujos de los papeles pintados o de las alfombras. El arte de fabricar tales cenefas y dibujos, el arte de la ornamentación, ha gozado de gran difusión entre la mayoría de las naciones desde la Antigüedad hasta el presente. En la figura 19 tenemos una muestra del dibujo de una techumbre egipcia que data de mediados del segundo milenio a. de J. C.

Al discutir los grupos de simetría de figuras finitas nos vimos obligados a distinguir entre los casos 1, 2, 3 y 5 cuando los grupos de

simetría no contienen giros de ángulo arbitrariamente pequeño, y entre los casos 4 y 6 cuando los grupos contienen tales giros. Al estudiar los grupos de simetría de figuras infinitas, especialmente en el caso tridimensional, esta división en grupos discretos y grupos de transformaciones arbitrariamente pequeñas es todavía más importante. Comenzaremos, por tanto, con una división más cuidadosa de estos casos.

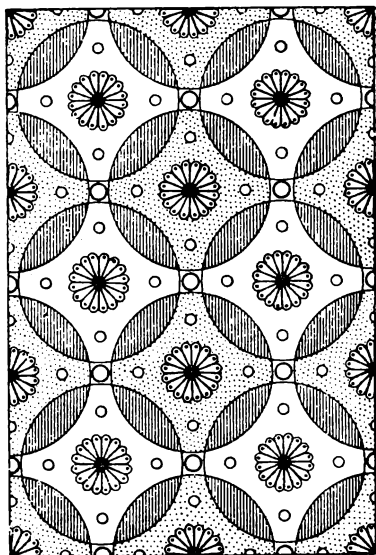


FIG. 19

fijo dicho punto. Este tipo contiene los grupos K_1 , K_2 , K_3 y K_6 de la lista anterior.

II. No existen puntos fijos en el plano, pero sí una recta que es transformada en sí misma bajo todas las transformaciones del grupo. Se dice que esta recta es un eje del grupo. Los grupos de simetría de este tipo aparecen en los ornamentos en forma de banda infinita (cenefa). En total, existen siete grupos de este tipo:

1. El grupo de simetría L_1 , que consta solamente de traslaciones cuyos módulos son múltiplos de un cierto segmento a .

2. El grupo L_2 , que se obtiene del L_1 añadiendo un giro de 180° de centro un punto cualquiera del eje del grupo.

Un grupo de movimientos del plano se llama *discreto* si todo punto del plano se puede encerrar en un círculo tal que todo movimiento del grupo, o bien deja el punto invariable, o bien lo transforma en un punto fuera del círculo elegido.

De la misma forma podemos determinar todos los grupos discretos de movimientos del plano. Todos estos grupos son grupos de simetría de figuras planas. Es natural distinguir tres tipos de grupos discretos de simetría:

- I. Existe un punto que permanece invariable bajo cualquier transformación de simetría; es decir, todas las simetrías dejan

3. El grupo L_3 , que se obtiene del L_2 añadiendo una simetría respecto a una recta perpendicular al eje del grupo.

4. El grupo L_4 , que se obtiene del L_1 añadiendo la simetría respecto al eje del grupo.

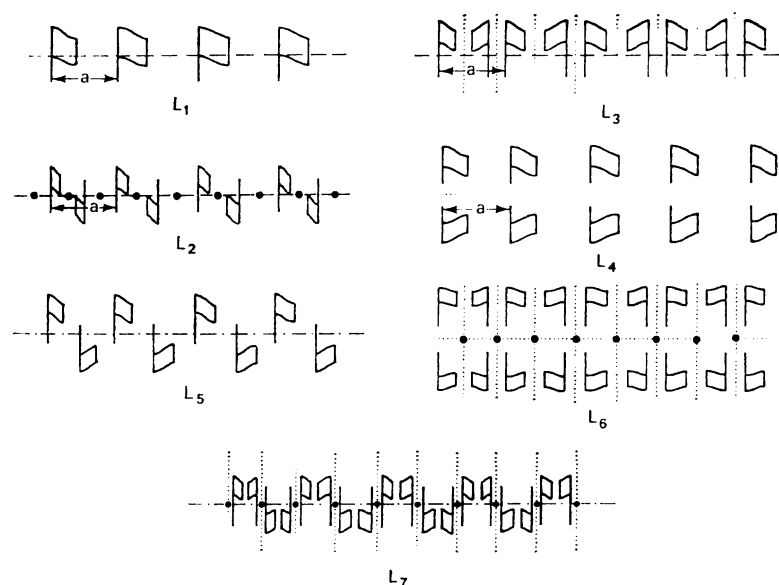
5. El grupo L_5 , que se obtiene añadiendo al grupo L_1 una traslación de módulo $a/2$ combinada con una simetría respecto al eje del grupo.

6. El grupo L_6 , que se obtiene añadiendo al grupo L_4 una simetría respecto a un eje perpendicular al eje del grupo.

7. El grupo L_7 , que se obtiene del L_5 añadiendo una simetría respecto a un eje perpendicular al eje del grupo.

La tabla 2 da ejemplos de «cenefas» correspondientes a cada uno de los grupos desde L_1 hasta L_7 .

TABLA 2



III. No existe ningún punto ni recta que se transforme en sí mismo bajo todas las transformaciones del grupo. Los grupos de este tipo se llaman *grupos planos de Fedorov*. Son los grupos de simetría de ornamentos planos infinitos. En total hay 17: cinco constan de

movimientos de primera especie solamente, y los doce restantes de movimientos de primera y segunda especie.

En la tabla 3 se dan ejemplos de ornamentos correspondientes a cada uno de los diecisiete grupos de Fedorov; cada grupo consta precisamente de aquellos movimientos que transforman una de las banderitas dibujadas en el diagrama en otra cualquiera del mismo diagrama.

Es interesante observar que los maestros en el arte de la ornamentación han descubierto en la práctica ornamentos con todos los posibles grupos de simetría; a la teoría de grupos le correspondió demostrar que no hay más formas posibles.

Grupos cristalográficos

En 1890, el eminente cristalógrafo y geómetra ruso E. S. Fedorov resolvió por métodos de la teoría de grupos uno de los problemas fundamentales de la cristalografía: clasificar los sistemas regulares de puntos en el espacio. Este fue el primer ejemplo de aplicación directa de la teoría de grupos a la solución de un problema importante de las ciencias naturales, causando un impacto sustancial en el desarrollo de la teoría de grupos.

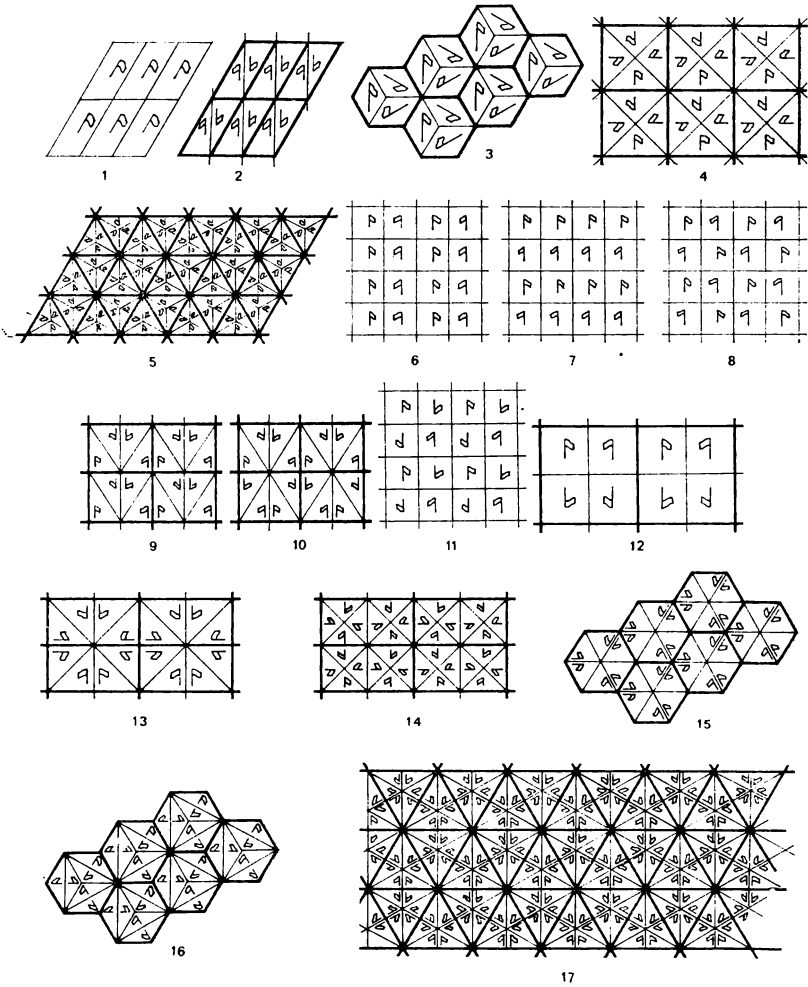
Un cristal tiene la peculiaridad de que sus átomos forman, en cierto sentido, un sistema regular en el espacio. Consideremos los movimientos espaciales que llevan los puntos de un sistema a puntos del mismo sistema. Estos movimientos forman un grupo cuyas propiedades permiten formular con mayor precisión el concepto de sistema regular de puntos.

Un sistema de puntos en el espacio se llama *sistema espacial regular de puntos* si:

1. todo punto del sistema se puede transportar a cualquier otro punto del sistema mediante movimientos que transforman el sistema en sí mismo;
2. ninguna esfera de radio finito contiene infinitos puntos del sistema;
3. existe un número positivo r tal que toda esfera de radio r contiene al menos uno de los puntos del sistema.

El problema de estudiar la estructura de los cristales está íntimamente ligada a la clasificación de los sistemas espaciales regulares de puntos, que a su vez está relacionada con la clasificación de los grupos discretos de movimientos en el espacio. Del mismo modo

TABLA 3



que en el plano, un grupo de movimientos H del espacio se llama *discreto* si para todo punto A existe una esfera de radio r y centro A tal que todo movimiento de H , o deja fijo el punto A , o lo lleva fuera de la esfera.

Se puede demostrar que el conjunto de los movimientos del espacio que llevan un sistema espacial regular de puntos a coincidir consigo mismo es necesariamente un grupo discreto y que todos los puntos del sistema se pueden obtener a partir de un punto dado del sistema transformando éste por todos los movimientos del grupo. Inversamente, dado un cierto grupo discreto H , si tomamos un punto arbitrario A y lo sometemos a todas las transformaciones del grupo obtenemos un sistema de puntos que cumple las propiedades 1 y 2. Añadiendo condiciones muy sencillas se pueden aislar de entre los grupos discretos aquellos que para puntos A convenientemente elegidos dan, de hecho, sistemas regulares de puntos, es decir, sistemas de puntos con las tres propiedades (1), (2), (3). Tales grupos discretos se llaman grupos *cristalográficos* o *de Fedorov*. De lo que llevamos dicho se desprende que el paso más importante en el estudio de los sistemas espaciales de puntos, y el primero, consiste en encontrar los grupos de Fedorov. Se ha comprobado que para los fines que persiguen las ciencias de la naturaleza es preciso considerar grupos que constan no sólo de movimientos propios, sino también aquellos que contienen movimientos propios e impropios, es decir, que incluyen simetrías respecto a ejes. El número de grupos de Fedorov formados solamente de movimientos propios es mucho más pequeño que el de los compuestos de movimientos propios e impropios; sólo en este último caso, más general, la variedad de sistemas espaciales de puntos regulares obtenida abarca toda la diversidad de estructuras de cristales que existen en la naturaleza.

Es interesante observar que, en contraposición al caso del plano tratado previamente, sólo la teoría de grupos permite analizar este número excepcionalmente grande de posibilidades.

La complejidad del problema en el espacio comparado con el del plano se ve claramente en la tabla 4.

TABLA 4

Número de grupos espaciales de Fedorov

Grupos que contienen sólo movimiento de primera especie.	65
Grupos que contienen también movimientos de segunda especie ..	165
Total.	230

Incluso hoy, una deducción y enumeración detallada de todos los grupos de Fedorov en el espacio requiere varias docenas de páginas de texto. Nos limitaremos, por tanto, a dar cuenta de los resultados cuantitativos y a remitir al lector interesado a la literatura especializada².

El desarrollo moderno de la cristalografía ha hecho necesario introducir una nueva extensión del concepto de simetría. Estas nuevas posibilidades metodológicas están tratadas en el libro sobre cristalografía del académico A. V. Subnikov, *Simetría y antisimetría de figuras finitas*, Akad. Nauk URSS, Moscú, 1951.

§5. Grupos de Galois

Los resultados desarrollados en las páginas precedentes dan una idea del papel que desempeña la teoría de grupos en la solución del problema de la clasificación de cristales. Sin embargo, este problema no fue el motivo de la creación de la teoría de grupos. Aproximadamente cien años antes, Lagrange observó una relación entre las propiedades de simetría de las raíces de una ecuación algebraica y la posibilidad de resolver la ecuación por radicales. Esta relación fue objeto de profundos estudios por parte de los famosos matemáticos Abel y Galois en los primeros treinta años del siglo XIX; y así llegaron a una solución del célebre problema de la resolubilidad de ecuaciones algebraicas mediante radicales. La solución se basaba por entero en una sutil investigación de las propiedades de los grupos de permutaciones y fue, de hecho, el comienzo de la teoría de grupos.

El estudio de las relaciones entre las propiedades de las ecuaciones algebraicas y las propiedades de los grupos constituye hoy día el objeto de una extensa teoría que se conoce con el nombre de teoría de Galois.

En el capítulo 4 dimos ya una idea de la historia del problema y de la importancia de la teoría de Galois. Sin embargo, puesto que la teoría de Galois desempeñó tan importante papel en el desarrollo de la teoría de grupos, repetiremos los hechos básicos de esta teoría,

² Una exposición detallada de los grupos de movimientos de primera especie discretos del plano se encuentra en el libro de D. Hilbert y S. E. Cohn-Vossen, *Geometry and the imagination* [Geometría e imaginación], Chelsea, Nueva York, 1952. Sobre los grupos cristalográficos en el espacio, el lector puede consultar el fundamental artículo de E. S. Fedorov, «Simetría de los sistemas regulares de figuras», *Obras completas*, Akad. Nauk URSS, Moscú, 1949.

pero en una forma más apropiada al objeto de arrojar nueva luz sobre la propia teoría de grupos. La demostración estos resultados requiere la utilización de otros muchos conceptos auxiliares, que omitiremos.

El grupo de una ecuación algebraica

Consideremos una ecuación de grado n :

$$x^n + a_1 x^{n-1} + \cdots + a_n = 0, \quad (6)$$

cuyos coeficientes se suponen conocidos; por ejemplo, ciertos números complejos. El conjunto de todas las cantidades que se pueden obtener a partir de los coeficientes de la ecuación mediante un número finito de operaciones de adición, sustracción, multiplicación y división se llama *cuerpo de base* o dominio de racionalidad de la ecuación.

Por ejemplo, si la ecuación tiene coeficientes racionales, el dominio de racionalidad está formado por todos los números racionales; si la ecuación tiene la forma $x^2 + \sqrt{2}x + 1 = 0$, entonces el dominio de racionalidad consiste en todos los números de la forma $a + b\sqrt{2}$, donde a, b son números racionales.

Denotaremos las raíces de esta ecuación por $\xi_1, \dots, \xi_n$. El conjunto de todas las cantidades obtenidas mediante un número finito de operaciones de adición, sustracción, multiplicación y división a partir de las raíces $\xi_1, \dots, \xi_n$, se llama *campo de descomposición* de la ecuación. Por ejemplo, el campo de descomposición de la ecuación $x^2 + 1 = 0$ es el conjunto de los números complejos $a + bi$ con a, b racionales; y el campo de descomposición de la anterior ecuación $x^2 + \sqrt{2}x + 1 = 0$ es el conjunto de los números de la forma $a + bi + c\sqrt{2} + di\sqrt{2}$, donde a, b, c, d son números racionales.

Por las fórmulas de Vieta, los coeficientes de la ecuación se obtienen a partir de las raíces por medio de las operaciones de suma y multiplicación, y por tanto el campo de descomposición de una ecuación contiene siempre a su cuerpo de base. A veces ambos coinciden.

Una aplicación uno a uno del campo de descomposición en sí mismo se dice que es un automorfismo del campo de descomposición respecto al cuerpo de base si a la suma y producto de dos elementos hace corresponder la suma y producto respectivamente de sus imágenes, y si transforma todo elemento del cuerpo de base en sí mismo. Podemos describir estas propiedades mediante las fórmulas

$$(a + b)A = aA + bA, \quad (ab)A = aA \cdot bA, \quad aA = a \\ (a, b \in K, \quad a \in P), \quad (7)$$

donde aA es la imagen de a ; esto es, aA es el elemento en que a se transforma mediante A ; P es el cuerpo de base; y K es el campo de descomposición.

Por el principio general expresado en §2, el conjunto de todos los automorfismos del campo de descomposición relativos al cuerpo de base es un grupo. Este grupo se llama *grupo de Galois* de la ecuación dada.

Para darnos una idea más concreta del grupo de Galois, observemos primeramente que todos los automorfismos del cuerpo de Galois transforman cada raíz de la ecuación en otra raíz. Sea x una raíz de la ecuación (6); operando en ambos miembros de la ecuación con el automorfismo A y haciendo uso de las propiedades (7) obtenemos

$$(xA)^n + a_1A(xA)^{n-1} + \cdots + a_nA = 0 \cdot A;$$

puesto que $0 \cdot A = 0$, $a_iA = a_i$, obtenemos

$$(xA)^n + a_1(xA)^{n-1} + \cdots + a_n = 0,$$

como queríamos probar. En consecuencia, cada automorfismo A efectúa una determinada permutación del conjunto de las raíces de la ecuación. Por otra parte, si conocemos esta permutación conocemos también el automorfismo, ya que todos los elementos del campo de descomposición se obtienen de las raíces mediante operaciones aritméticas solamente. Esto demuestra que en vez del grupo de los automorfismos podemos considerar el de las permutaciones de las raíces de la ecuación correspondiente. De aquí se deduce, en particular, que todos los grupos de Galois son finitos.

Hallar el grupo de Galois de una ecuación suele ser complicado, y sólo en algunos casos es relativamente fácil. Consideremos, por ejemplo, la ecuación (6) con coeficientes literales $a_1, a_2, \cdots, a_n$. El cuerpo de base de esta ecuación está formado por las fracciones racionales de sus coeficientes, es decir, las fracciones cuyos numeradores y denominadores son polinomios en $a_1, \cdots, a_n$. El campo de descomposición está formado por fracciones racionales de las raíces de la ecuación $\xi_1, \cdots, \xi_n$, que están relacionadas con los coeficientes por las fórmulas

$$\begin{aligned} -a_1 &= \xi_1 + \xi_2 + \cdots + \xi_n, \\ a_2 &= \xi_1\xi_2 + \xi_1\xi_3 + \cdots + \xi_{n-1}\xi_n, \\ &\dots\dots\dots \\ (-1)^na_n &= \xi_1\xi_2 \cdots \xi_n. \end{aligned} \tag{8}$$

Puesto que la ecuación (6) es «general», podemos considerar sus raíces como variables independientes. Entonces toda permutación de estas raíces da lugar a un automorfismo del campo de descomposición. Las fórmulas (8) muestran que cualquier automorfismo transforma los coeficientes en ellos mismos y, además, todas las fracciones formadas con los coeficientes también en ellas mismas. Así, pues, el grupo de Galois de la ecuación general de grado n es esencialmente el grupo simétrico de todas las permutaciones de n letras.

Podemos también indicar ecuaciones con coeficientes numéricos que tienen como grupo de Galois el grupo simétrico. Por ejemplo, se ha demostrado que el grupo de Galois de la ecuación

$$1 - \frac{n}{1}x + \frac{n(n-1)}{1 \cdot 2} \cdot \frac{1}{1 \cdot 2}x^2 - \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} \cdot \frac{1}{1 \cdot 2 \cdot 3}x^3 + \\ + \cdots + (-1)^n \frac{1}{1 \cdot 2 \cdots n} x^n = 0 \quad (9)$$

para un n arbitrario es el grupo simétrico de grado n .

Se conocen métodos generales para construir ecuaciones que tengan grupos de Galois determinados de antemano, pero a condición de que puedan tomarse coeficientes arbitrarios. Sin embargo, si se exige que los coeficientes de la ecuación sean racionales, hasta el momento sólo se conocen métodos de construcción para tipos particulares de grupos. Notables progresos en esta dirección se deben al matemático soviético I. R. Safarevic, quien ha encontrado métodos para construir ecuaciones con coeficientes reales que tengan como grupo de Galois un grupo resoluble cualquiera. Sin embargo, este problema no está resuelto en general.

Resolubilidad de ecuaciones mediante radicales

El grupo de Galois caracteriza, como se deduce de su definición, la simetría intrínseca de las raíces de la ecuación. Todos los problemas fundamentales relativos a la posibilidad de reducir la solución de una ecuación a la de ecuaciones de grados inferiores y otros muchos se pueden formular como problemas sobre la estructura del grupo de Galois; y el grupo de Galois de toda ecuación de grado n es un cierto grupo de permutaciones de grado n , esto es, un objeto finito, en el cual todas las relaciones, al menos teóricamente, se pueden encontrar por métodos de tanteo.

El estudio del grupo de Galois es un valioso método para la resolución de problemas relativos a ecuaciones algebraicas de grado superior. Por ejemplo, se puede demostrar que una ecuación es resoluble por radicales si, y sólo si, su grupo de Galois es resoluble (para la definición de grupo resoluble véase §3). Hemos mencionado ya que los grupos simétricos de grado 2, 3 y 4 son resolubles por radicales. Los grupos de Galois de las ecuaciones «generales» de grado 5, 6, etcétera, son los grupos simétricos del mismo grado. Pero estos grupos no son resolubles. Se sigue de aquí que las ecuaciones generales de grados superiores a 4 no se pueden resolver por radicales.

Entre las ecuaciones que no se pueden resolver por radicales están también las ecuaciones (9) para $n > 4$, ya que su grupo de Galois es el grupo simétrico.

§6. *Conceptos fundamentales de la teoría general de grupos*

La teoría de grupos surgió en el siglo XIX fundamentalmente como la teoría de grupos de transformaciones. Sin embargo, a lo largo del tiempo fue haciéndose cada vez más claro que de entre los resultados obtenidos, los más importantes dependen solamente del hecho de que las transformaciones se pueden multiplicar y de que esta operación tiene un cierto número de propiedades características. Por otra parte, se encontraron objetos sin ninguna relación con las transformaciones, con los que se podía efectuar una cierta operación (que por ahora llamaremos multiplicación) con las mismas propiedades que en los grupos de transformaciones y a los que se podían aplicar los principales resultados de la teoría de grupos de transformaciones. Como consecuencia de ello, a finales del siglo pasado el concepto de grupo fue aplicado no sólo a sistemas de transformaciones, sino también a sistemas de elementos arbitrarios.

Definición general de grupo

La siguiente definición de grupo se acepta con carácter general en nuestros días: supongamos que a cada par de elementos a, b , tomados en un orden definido, de un conjunto arbitrario G se le asocia otro elemento c bien definido del conjunto G . Se dice entonces que sobre G se ha dado una operación. Es habitual introducir nombres especiales para las operaciones: suma, multiplicación, composición. El elemento de G que corresponde al par a, b se llama entonces suma, producto, y composición de los elementos a, b , y se denota

por $a + b$, ab , $a * b$, respectivamente, los nombres de «suma» y «multiplicación» se usan incluso en casos en los que la operación en cuestión no tiene nada que ver con las operaciones ordinarias de suma y multiplicación de números.

Un conjunto G junto con una operación $*$ definida sobre él se llama *grupo* con respecto a esta operación si se cumplen los siguientes axiomas:

1. Para cada tres elementos x, y, z de G

$$x * (y * z) = (x * y) * z \quad (\text{ley asociativa}).$$

2. Entre los elementos de G existe un elemento e tal que para todo x de G

$$x * e = e * x = x.$$

3. Para todo elemento a de G existe un elemento a^{-1} de G tal que

$$a * a^{-1} = a^{-1} * a = e.$$

El elemento e descrito en el axioma 2 recibe el nombre de elemento neutro del grupo, y el elemento a^{-1} cuya existencia se postula en el axioma 3, se llama inverso de a . Si la operación del grupo se llama adición o multiplicación, el elemento neutro se llama cero o unidad, respectivamente, y los axiomas de grupo toman la forma

$$\begin{array}{ll} (1) \ x + (y + z) = (x + y) + z, & (1) \ x(yz) = (xy)z, \\ (2) \ x + 0 = 0 + x = x, & (2) \ xe = ex = x, \\ (3) \ x + (-x) = (-x) + x = 0, & (3) \ xx^{-1} = x^{-1}x = e. \end{array}$$

En las secciones precedentes hemos discutido muchos ejemplos de grupos. Los elementos de estos grupos eran transformaciones, y la operación del grupo, la multiplicación de transformaciones. El conjunto de los números $0, \pm 1, \pm 2, \dots$ forma también un grupo para la operación de adición, ya que la suma de enteros es otro entero y además la operación es asociativa; el elemento neutro es el entero 0 y para todo número a de nuestro conjunto existe el opuesto $-a$. Otro ejemplo de grupo es el conjunto de los números reales (excepto el cero) para la multiplicación. El producto de dos números reales es otro número real; la operación es asociativa; el elemento neutro es el número 1; y todo número real distinto de cero tiene un inverso

$a^{-1} = 1/a$. El número de ejemplos similares se podría aumentar indefinidamente.

Aunque las operaciones del grupo pueden recibir diferentes nombres, convenimos de ahora en adelante en llamarlas, casi siempre, multiplicación. Los conceptos de subgrupo, de potencias de un elemento, de grupo cíclico y de orden de un elemento de un grupo se definen exactamente igual que para los grupos de transformaciones, y no las repetimos (ver §3). Solamente mencionaremos que un elemento a de un grupo se llama conjugado de otro b si existe un elemento x en G tal que $b = x^{-1}ax$. Puesto que $a = a^{-1}aa$, todo elemento de un grupo es conjugado de sí mismo. Además, de $b = x^{-1}ax$ se sigue inmediatamente que $a = bxx^{-1}$ o bien $a = (x^{-1})^{-1}bx^{-1}$, esto es, si a es conjugado de b , entonces b es conjugado de a . Finalmente, si $b = x^{-1}ax$ y $c = y^{-1}by$, entonces

$$c = y^{-1}x^{-1}axy = (xy)^{-1}a(xy).$$

Por tanto, dos elementos conjugados con un tercero son conjugados entre sí. Estas propiedades muestran que los elementos de un grupo se dividen en clases disjuntas de elementos conjugados. Si el grupo es conmutativo, es decir, si se verifica $xy = yx$ para todo x e y , entonces los elementos conjugados coinciden y cada clase consta de un solo elemento.

Isomorfismo

Cabe distinguir dos aspectos en el concepto de grupo. Para dar un grupo debemos: 1) indicar qué objetos son sus elementos, y 2) indicar la ley de multiplicación de los elementos. De acuerdo con esto, el estudio de las propiedades de un grupo se puede realizar desde distintos puntos de vista. Se pueden estudiar las relaciones entre las propiedades individuales de elementos del grupo y de conjuntos de ellos, o bien sus propiedades en relación con la operación del grupo. Este punto de vista se adopta a menudo en el estudio de grupos concretos; por ejemplo, el grupo de los movimientos del espacio del plano. Sin embargo, podemos también estudiar aquellas propiedades del grupo que se pueden expresar por completo en función de las propiedades de la operación del grupo. Este punto de vista es característico de la teoría *general* o *abstracta* de grupos. Se puede expresar más claramente por medio del concepto de *isomorfismo*.

Se dice que dos grupos son *isomorfos* si los elementos de uno de

ellos se pueden asociar con los del otro de modo que al producto de dos elementos arbitrarios del primero corresponda el producto de los correspondientes elementos del segundo. Si una correspondencia uno a uno tiene esta propiedad se dice que es un *isomorfismo*.

Es fácil darse cuenta de que los elementos que se corresponden en un isomorfismo tienen idénticas propiedades respecto a la operación del grupo. Así, un isomorfismo transforma el elemento neutro, elementos inversos, elementos de orden n , y sus grupos de un grupo en el elemento neutro, elementos inversos, elementos del mismo orden, y subgrupos, respectivamente, del segundo grupo. Se puede decir, por tanto, que la teoría abstracta de grupos estudia solamente las propiedades del grupo que se conservan en las aplicaciones isomorfas. Por ejemplo, desde el punto de vista de la teoría abstracta de grupos, el grupo de todas las permutaciones de cuatro elementos y el grupo de los movimientos propios e impropios del espacio que transforman en sí mismo un tetraedro regular tienen propiedades idénticas porque son isomorfos. En efecto, los movimientos en cuestión llevan los vértices del tetraedro de nuevo sobre los vértices. El número de estos movimientos es 24. Asociando a cada movimiento la permutación de vértices que origina, obtenemos una correspondencia uno a uno entre los elementos de los dos grupos que es el isomorfismo buscado.

Un ejemplo notable de isomorfismos lo da la teoría de logaritmos. Asociando a cada número positivo su logaritmo, obtenemos una aplicación uno a uno entre los números reales positivos y todos los números reales. La relación $\log(xy) = \log x + \log y$ muestra que esta correspondencia es un isomorfismo entre el grupo de los números reales positivos con la operación de multiplicar y el grupo de todos los reales con la suma. La importancia práctica de este isomorfismo es bien conocida.

Los grupos finitos de distintos órdenes son ejemplos de grupos no isomorfos. Como ya hemos mencionado, un grupo abstracto está determinado por la ley de multiplicación de sus elementos, independientemente de su naturaleza, de modo que grupos concretos distintos pero isomorfos se pueden considerar como modelos del mismo grupo abstracto.

Un grupo abstracto se puede dar por distintos métodos, de los cuales el más natural —al menos para grupos finitos— es dar una «tabla de multiplicación».

Para un grupo de orden n cuyos elementos están escritos en un orden arbitrario, la tabla de multiplicación consiste en un cuadrado

dividido en n filas y n columnas. En la intersección de la fila i con la columna j se escribe el elemento producto del elemento que lleva el número i por el que lleva el número j . Este tipo de tabla para grupos finitos se llama a veces cuadrado de Cayley.

Sin embargo, en la práctica casi nunca es conveniente dar un grupo mediante una tabla porque es una tarea muy pesada.

Existen otros métodos de dar un grupo abstracto. Uno de ellos —el que hace uso de generadores y relaciones de definición— lo hemos visto ya. Sin embargo, lo más frecuente es definir un grupo abstracto dando un grupo concreto isomorfo a él, en particular, un grupo de transformaciones.

Naturalmente, se plantea ahora el problema de si todo grupo abstracto se puede considerar como un grupo de transformaciones. El siguiente teorema da la respuesta: todo grupo abstracto es isomorfo a un grupo de transformaciones del conjunto de sus elementos.

Probémoslo. Sea g un elemento fijo de G . Denotemos por A_g la transformación que asocia a cada elemento x de G el elemento xg . La transformación es uno a uno, ya que la ecuación

$$xA_g = xg = a$$

tiene para cada a una única solución $x = ag^{-1}$. Por otra parte, el producto gh se asocia con el producto de las correspondientes transformaciones $A_g A_h$, ya que

$$xA_{gh} = x(gh) = (xg)h = (xA_g)A_h = x(A_g A_h).$$

Al elemento neutro e de G corresponde la identidad, y al elemento g^{-1} , la transformación inversa. Por tanto, el conjunto I de las transformaciones correspondientes a los elementos de G es un grupo de transformaciones isomorfo a G . Es fácil verificar que, si el número de elementos de G es mayor que 2, el conjunto I no agota todas las transformaciones de G y es solamente un subgrupo del grupo «simétrico» de todas las transformaciones posibles del conjunto.

Subgrupos normales y grupos factor

Sean P y Q colecciones arbitrarias de un grupo G . El producto PQ , producto de P por Q , es el nombre que se da al conjunto de los elementos de G que se pueden representar como producto de un elemento de P por uno de Q . En particular, el producto gP , donde g es un elemento de G , es el conjunto de productos de g por cada elemento del conjunto P .

El subgrupo H de G se llama *subgrupo normal* o *invariante* de G , $gH = Hg$, para todo g de G . Los conjuntos de la forma gH y Hg , donde H es un subgrupo arbitrario, se llaman respectivamente *clase a la derecha* y *a la izquierda* de H en G . Podemos decir, por tanto, que los subgrupos normales están completamente caracterizados por la propiedad de que para ellos las clases a la derecha y a la izquierda correspondientes al mismo elemento coinciden.

Si H es un subgrupo normal, el producto de dos clases es otra clase, como es fácil de ver: $aH \cdot bH = ab \cdot H$. El propio subgrupo H es una clase correspondiente al elemento unidad o a uno cualquiera de sus elementos, ya que si h pertenece a H , $hH = H$. El producto de clases es asociativo

$$(aH \cdot bH)cH = (ab \cdot c)H = (a \cdot bc)H = aH(bH \cdot cH).$$

El subgrupo H juega el papel de elemento neutro en esta multiplicación: $H \cdot aH = eH \cdot aH = (ea)H = aH$; análogamente, $aH \cdot H = aH$. La clase $a^{-1}H$ es la inversa de aH , ya que $aH \cdot a^{-1}H = aa^{-1}H = H$. Considerando, por tanto, cada clase de un subgrupo normal como un elemento de un nuevo conjunto, vemos que este conjunto es un grupo para la multiplicación de clases. Este grupo recibe el nombre de *grupo factor* (o *grupo cociente*) de G respecto al subgrupo normal H y se denota por G/H .

Es fácil demostrar que para grupos finitos toda clase de un subgrupo arbitrario contiene el mismo número de elementos que dicho subgrupo, y que clases distintas no tienen ningún elemento en común. Se sigue de aquí que el número de clases de un cierto subgrupo en un grupo finito es igual al orden de G dividido por el de H ; y esto implica el importante teorema de Lagrange que afirma que el orden de todo subgrupo de un grupo finito es un divisor del orden del grupo.

De la definición de subgrupo normal se deduce que todo subgrupo de un grupo abeliano es normal. El otro caso extremo lo constituyen los grupos llamados *simples*, en los cuales ningún subgrupo, salvo el subgrupo unidad y el propio grupo, es normal. Aparte de los grupos abelianos y los grupos simples, los grupos resolubles definidos en la sección 3 son también muy importantes. Se puede demostrar que los grupos resolubles tienen una cadena finita $G, G_1, \dots, G_k$ de subgrupos normales, el primero de los cuales coincide con G , cada grupo está contenido en su predecesor, el último es la unidad y todos los grupos factores $G/G_1, G/G_2, \dots, G_{k-1}/G_k$ son abelianos.

Homomorfismos

El concepto de grupo factor está íntimamente relacionado con el de homomorfismo, que es fundamental en toda la teoría de grupos.

Una aplicación unívoca del conjunto de elementos de un grupo G sobre los de un grupo H se llama *homomorfismo* o *aplicación homomorfa* si el producto de cada dos elementos del primer grupo se aplica sobre el producto de los elementos correspondientes del segundo.

Si denotamos por x' el elemento de H correspondiente a un elemento x de G , un homomorfismo se caracteriza por la siguiente ecuación:

$$(x_1 x_2)' = x'_1 x'_2.$$

De las definiciones de homomorfismo e isomorfismo se deduce que un isomorfismo es siempre una aplicación uno a uno, mientras que un homomorfismo es unívoco sólo en una dirección: a cada elemento de G corresponde un único elemento de H , pero distintos elementos de G tienen o pueden tener la misma imagen en H . Se puede decir, en cierto sentido, que si entre G y H existe un isomorfismo, H es una copia exacta de G ; pero si la aplicación de G en H es un homomorfismo, entonces elementos distintos de G se pueden fundir en uno solo de H . Sin embargo, esta naturaleza «grosera» del homomorfismo no supone una deficiencia; por el contrario, es una gran ventaja, pues permite usar las aplicaciones homomorfas como un poderoso instrumento en el estudio de las propiedades de grupos.

Los homomorfismos aparecen en muchas situaciones relacionadas con las transformaciones. Por ejemplo, consideremos el grupo de simetría del tetraedro regular (Fig. 20). Este grupo es isomorfo al grupo simétrico de las permutaciones de cuatro elementos, ya que existe uno y un solo movimiento (de primera o segunda especie) que lleva los vértices A_1, A_2, A_3, A_4 sobre otra permutación de ellos.

Consideremos las rectas l_1, l_2, l_3 que unen los puntos medios de aristas opuestas. Todo movimiento que lleva el tetraedro a coincidir

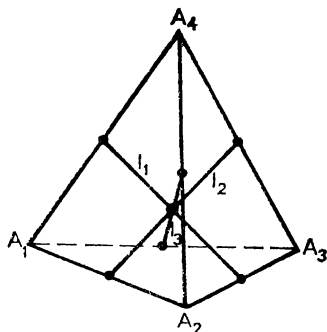


FIG. 20

consigo mismo da lugar a una permutación de l_1, l_2, l_3 , y toda permutación de estas tres rectas es generada por una simetría del tetraedro. Evidentemente, al producto de transformaciones del tetraedro corresponde el producto de permutaciones de las rectas l_1, l_2, l_3 . De la figura 20 se deduce cómo se puede realizar de forma natural un homomorfismo entre el grupo simétrico de las permutaciones de los cuatro elementos A_1, A_2, A_3, A_4 y el grupo simétrico de las permutaciones de los tres elementos l_1, l_2, l_3 . No es difícil encontrar los elementos del grupo «mayor» que se «funden» en este homomorfismo.

Discutamos algunos ejemplos más. El conjunto de todas las permutaciones de n elementos para $n > 2$ es un grupo no conmutativo. Por otra parte, los números -1 y $+1$ también forman un grupo para la multiplicación. Asociemos ahora a cada permutación par o impar los números $+1$ y -1 respectivamente. Esta aplicación es un homomorfismo del grupo simétrico de las permutaciones de n elementos sobre el grupo $\{+1, -1\}$, ya que de acuerdo con el §3 el producto de permutaciones de igual paridad es una permutación par, y el producto de distinta paridad es impar.

Otro ejemplo: si asociamos a cada número real $x \neq 0$ su valor absoluto $|x|$, la aplicación resultante es un homomorfismo del grupo de todos los números reales para la multiplicación (excluido el cero) sobre el grupo de los reales positivos también para la multiplicación, ya que $|xy| = |x| |y|$.

Hemos mencionado ya que en el plano todo movimiento de primera especie A se puede escribir como un producto de un giro V_A de centro un punto fijo O por una cierta traslación D_A . Los giros de centro O forman un grupo. Por tanto, la correspondencia $A \rightarrow V_A$ aplica unívocamente el grupo de los movimientos del plano de primera especie en el grupo de giros del plano de centro el punto O . Probemos que esta aplicación es un homomorfismo. De la descomposición $A = V_A D_A, B = V_B D_B$ se sigue que

$$AB = V_A D_A V_B D_B = (V_A V_B) (V_B^{-1} D_A V_B D_B).$$

El primer paréntesis es un giro de centro O y el segundo es el producto de la traslación transformada $V_B^{-1} D_A V_B$ por la traslación D_B y es, por tanto, también una traslación. Esto demuestra que el producto de movimientos AB se asocia al producto de los correspondientes giros $V_A V_B$, es decir, que la aplicación en cuestión es un homomorfismo.

Probemos, por último, que el grupo factor G/N de un grupo

arbitrario G respecto a un subgrupo normal N es una imagen homomorfa de G .

Para ello, asociando a cada elemento g de G la clase gN que contiene a g , obtenemos el homomorfismo buscado de G en G/N , ya que al producto gh corresponde la clase ghN , que es el producto de las clases gN y hN correspondiente a los elementos g y h .

Pasando ahora a las propiedades generales de los homomorfismos, probemos que el elemento neutro se aplica en el elemento neutro y los elementos inversos en elementos inversos.

Para ello, si e es el elemento neutro de G , y e' , su imagen en H , se sigue entonces de $ee = e$ que $e'e' = e'$, de modo que si denotamos por ϵ el elemento neutro de H , obtenemos: $e' = e'e'^{-1} = \epsilon$. Esto prueba la primera afirmación. Sean ahora x e y elementos inversos de G , y x' e y' , sus imágenes en H . De $xy = e$ se sigue que $x'y' = e' = \epsilon$, es decir, que x' e y' son elementos inversos en H y por tanto que

$$(x^{-1})' = x'^{-1}.$$

Los hechos probados permiten hallar fácilmente la imagen de un producto arbitrario de elementos de G . Por ejemplo,

$$(ab^{-1}c^{-1}dh^{-1})' = a'(b^{-1})'(c^{-1})'d'(h^{-1})' = a'b'^{-1}c'^{-1}d'h'^{-1}.$$

El siguiente teorema es fundamental en toda la teoría de homomorfismos.

El conjunto N de elementos de un grupo G que se aplican mediante un homomorfismo en el elemento neutro e' de un grupo H es un subgrupo normal de G ; el conjunto de elementos de G que se aplican en un elemento fijo de H es una clase de N en G , y la correspondencia uno a uno que se establece así entre las clases de N en G y los elementos de H es un isomorfismo entre H y el grupo factor G/N .

Probemos el teorema. Sean a y b elementos arbitrarios de N . Esto significa que $a' = b' = e'$, donde, como antes, las primas denotan las imágenes elementos de G en H . Pero entonces

$$\begin{aligned}(ab)' &= a'b' = e'e' = e', \\ (a^{-1})' &= a'^{-1} = e'^{-1} = e';\end{aligned}$$

esto es, ab y los elementos inversos a^{-1} y b^{-1} pertenecen a N , que es por tanto un grupo. Además, para cualquier elemento g de G tenemos

$$(g^{-1}ag)' = g'^{-1}a'g' = g'^{-1}e'g' = g'^{-1}g' = e';$$

esto es, $g^{-1}ag$ pertenece a N para todo elemento de G y todo a de N , y de aquí se sigue que N es un subgrupo normal. Esto prueba la primera parte del teorema.

Para probar la segunda tomamos en G un elemento cualquiera g y consideramos el conjunto U de todos los elementos u de G , tales que su imagen u' coincide con g' , imagen de g . Supongamos que $u \in gN$, es decir, $u = gn$ donde $n \in N$; entonces $u' = g'n' = g'e' = g'$. Por tanto $gN \subset U$. Inversamente, si $u' = g'$, entonces $(g^{-1}u)' = g'^{-1}u' = g'^{-1}g' = e'$; esto es, $g^{-1}u = n$, donde n es un elemento de N . Por tanto $u = gn$, y así $U \subset gN$. De $gN \subset U$ y $U \subset gN$ se sigue que $U = gN$.

Por último, la tercera afirmación del teorema es obvia: a clases arbitrarias gN , hN del grupo factor G/N corresponden en H los elementos g' y h' , y el producto de las clases, por la fórmula

$$gN \cdot hN = ghN,$$

corresponde $(gh)' = g'h'$, como se requería.

El teorema sobre homomorfismos muestra que toda imagen homomorfa H de un grupo G es isomorfa al correspondiente grupo factor G/H . Así, pues, salvo un isomorfismo, todas las imágenes homomorfas de un grupo dado G se agotan en sus distintos grupos factores.

§7. Grupos continuos

Grupos de Lie; grupos continuos de transformaciones

Los progresos hechos mediante la teoría de grupos en la resolución de ecuaciones algebraicas de grado superior indujeron a los matemáticos de mediados del siglo pasado a intentar utilizar la teoría de grupos en la resolución de ecuaciones de otra forma, y en primer lugar de ecuaciones diferenciales, que tan importante papel juegan en las aplicaciones de las matemáticas. Este intento se vio coronado por el éxito. Aunque el lugar que ocupan los grupos en ecuaciones diferenciales es distinto del que ocupan en la teoría de ecuaciones algebraicas, los estudios sobre la aplicación de la teoría de grupos a la resolución de ecuaciones diferenciales condujeron a una extensión sustancial del propio concepto de grupo y a la creación de la nueva teoría de los llamados grupos continuos y grupos de Lie, que han demostrado ser muy importantes para el desarrollo de las más diversas ramas de las matemáticas.

$$\begin{aligned} z_1 &= \phi_1(x_1, x_2, \dots, x_r; y_1, y_2, \dots, y_r), \\ z_2 &= \phi_2(x_1, x_2, \dots, x_r; y_1, y_2, \dots, y_r), \\ &\dots\dots\dots \\ z_r &= \phi_r(x_1, x_2, \dots, x_r; y_1, y_2, \dots, y_r). \end{aligned}$$

Consideremos una transformación uno a uno del conjunto de los puntos de un espacio geométrico arbitrario que no altere las relaciones básicas entre las figuras que se estudian en esa geometría. La colección de estas transformaciones forma un grupo, que habitual-

mente se llama grupo de los movimientos o de los automorfismos de dicha geometría. El grupo de movimientos caracteriza completamente la geometría dada, ya que, conocido dicho grupo, se puede considerar la correspondiente geometría como el estudio de aquellas propiedades de la colección de puntos que son conservadas por las transformaciones del grupo. El método de clasificar los distintos sistemas geométricos por sus grupos de movimientos fue introducido en la segunda mitad del pasado siglo por el matemático F. Klein. Este método y los distintos sistemas geométricos han sido tratados en el capítulo 16. Mencionaremos solamente que los grupos de movimientos de todos los sistemas geométricos estudiados en el siglo pasado eran grupos de Lie. En vista de esto, la tarea de estudiar los grupos de Lie cobró especial importancia.

Debido a sus múltiples conexiones con los más diversos dominios de las matemáticas y la mecánica, la teoría de los grupos de Lie se ha desarrollado rápidamente desde su fundación hasta nuestros días. Y así sucede que ciertos problemas no resueltos aún para grupos finitos se resolvieron de forma relativamente rápida para grupos de Lie. Por ejemplo, poco se ha avanzado hasta ahora en la clasificación de los grupos finitos simples (es decir, grupos finitos que no tienen subgrupos normales no triviales); en cambio, la correspondiente clasificación de los grupos de Lie simples fue obtenida por Killing y Cartan ya a finales del siglo pasado. Mediante el desarrollo de la teoría de los grupos de Lie, los matemáticos soviéticos V. V. Morozov, A. I. Malsev y E. B. Dynkin han encontrado una solución completa del importante e interesante problema de clasificar los *subgrupos* simples de los grupos de Lie. La teoría de los grupos de Lie fue desarrollada en otra dirección por los matemáticos soviéticos I. M. Gelfand y M. A. Naimark, que han encontrado lo que se denomina representaciones continuas de los grupos simples de Lie mediante transformaciones unitarias de un espacio de Hilbert; este último trabajo es de particular interés para el análisis y la física.

El estudio de los grupos de Lie procede por medio del peculiar aparato de los llamados «grupos infinitesimales» o álgebra de Lie, que serán discutidos con más detalle en el §3.

Grupos topológicos

Junto a una amplia extensión de la teoría clásica de los grupos de Lie, en la URSS se realizaron excepcionales avances en la teoría, más general, de los grupos topológicos o continuos. En contraste con el concepto de grupo de Lie, donde se requiere que los elementos

de un grupo estén definidos mediante un sistema finito de parámetros y que la multiplicación pueda ser expresada mediante funciones diferenciables, el concepto de grupo topológico es más simple y más amplio. Un grupo se llama *topológico* si además de la operación de grupo se define un concepto de proximidad entre sus elementos y si la proximidad de elementos del grupo implica la de sus productos y la de sus inversos.

En origen, el concepto de grupo topológico resultó necesario para poner orden en muchos de los conceptos fundamentales de la teoría de los grupos de Lie. Pero más tarde se vio también la gran importancia de este concepto para muchas otras ramas de las matemáticas. Los primeros trabajos sobre la teoría general de grupos topológicos se realizaron en los primeros años de la década de los veinte de nuestro siglo, pero los resultados fundamentales que permiten hablar de la creación de una nueva disciplina no fueron hallados hasta finales de los años veinte y comienzos de los treinta. Gran parte de ellos se deben al matemático L. S. Pontriagin, que sin lugar a dudas puede ser considerado como uno de los fundadores de la moderna teoría de grupos continuos. Su libro *Grupos topológicos*, que fue el primero en dar una exposición detallada de la teoría de grupos continuos, sigue siendo, después de veinte años, un texto básico en este campo.

§8. Grupos fundamentales

En todos los ejemplos concretos discutidos en las secciones precedentes, los grupos han aparecido casi siempre como grupos de transformaciones de uno u otro conjunto. Las únicas excepciones han sido los grupos de los números respecto a la suma y la multiplicación. Analicemos ahora un importante ejemplo que surge no como un grupo de transformaciones, sino como un cierto sistema algebraico con una operación.

El grupo fundamental

Consideremos una superficie S y sobre ella un punto M móvil. Si M se mueve sobre la superficie a lo largo de una curva continua que une A con B , obtenemos un camino de A a B . Este camino puede cortarse a sí mismo e incluso retroceder sobre sí mismo en algunas partes. Para especificar un camino no basta con dar la curva sobre la que el punto se mueve. Hay que indicar también las partes que el punto recorre más de una vez, así como la dirección de recorrido.

Por ejemplo, un punto puede recorrer un mismo círculo un número diferente de veces y en distintas direcciones, debiéndose considerar estos caminos como diferentes. Dos caminos con el mismo comienzo y el mismo final se llaman equivalentes si uno de ellos se puede transformar en el otro mediante un cambio continuo. En el plano o en la esfera dos caminos cualesquiera que unen el punto A con el B son equivalentes (Fig. 21). Sin embargo, sobre la superficie del toro, por ejemplo, los caminos cerrados U y V (Fig. 22) que comienzan y terminan en A no son equivalentes.

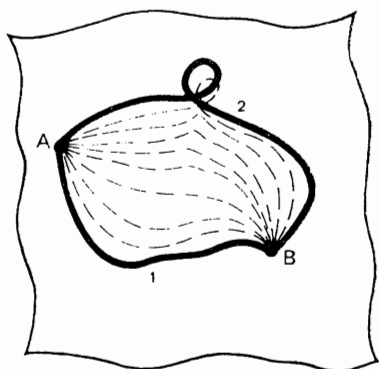


FIG. 21

Si en lugar de un toro consideramos un cilindro circular infinito extendido en ambas direcciones y tomamos sobre él un camino X (Fig. 23), es fácil comprender que

todo camino cerrado que comienza en A es equivalente a un camino de la forma X^n ($n = 0, \pm 1, \pm 2, \dots$), donde por X^n entendemos el camino X repetido n veces ($n > 0$); por X^0 , el camino reducido al

único punto A ; y por X^{-n} , el camino X^n recorrido en la dirección opuesta; por ejemplo, $Z \sim X^{-1}$, $Y \sim X^2$, $U \sim X^0$ (Fig. 23). Este ejemplo muestra la importancia del concepto de equivalencia de caminos: aunque existen infinitud de caminos cerrados distintos sobre

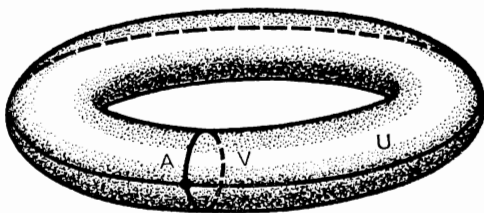


FIG. 22

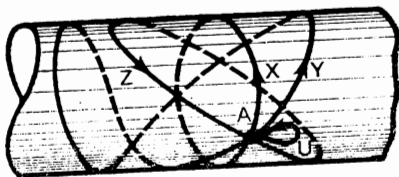


FIG. 23

único punto A ; y por X^{-n} , el camino X^n recorrido en la dirección opuesta; por ejemplo, $Z \sim X^{-1}$, $Y \sim X^2$, $U \sim X^0$ (Fig. 23). Este ejemplo muestra la importancia del concepto de equivalencia de caminos: aunque existen infinitud de caminos cerrados distintos sobre

el cilindro, todos ellos se reducen, salvo equivalencias, al círculo X recorrido en una u otra dirección un número suficiente de veces. Para $m \neq n$ los caminos X^m y X^n no son equivalentes.

Volviendo ahora a la discusión de una superficie arbitraria, supongamos dados dos caminos sobre ella: un camino U que va del punto A al B , y otro V , del punto B al C . Entonces, haciendo que un punto recorra primero el camino AB y a continuación el BC obtenemos un camino AC que de modo natural podemos llamar producto de los caminos $U = AB$ y $V = BC$ y que denotamos por UV . Si los caminos U, V son equivalentes a los caminos U_1, V_1 , respectivamente, entonces sus productos UV y U_1V_1 son también equivalentes. El producto de caminos es asociativo en el sentido de que si uno de los caminos $U(VW)$ está definido, también lo está el $(UV)W$, y los dos productos representan caminos equivalentes. Si dejamos que el punto M recorra el camino $U = AB$ pero en sentido opuesto, obtenemos el camino inverso $U^{-1} = BA$ que lleva de B a A . El producto del camino AB por su inverso BA es un camino cerrado equivalente al camino cero que consta del único punto A .

De acuerdo con la definición, no podemos multiplicar dos caminos más que cuando el final del primero coincide con el comienzo del segundo. Esta dificultad desaparece si consideramos caminos cerrados que parten del mismo punto inicial A . Podemos, entonces, multiplicar dos caminos tales, obteniendo otra vez un camino con el mismo punto inicial A . Además, para todo camino cerrado con el punto inicial en A , su camino inverso tiene las mismas propiedades.

Convenimos ahora en considerar los caminos equivalentes como representaciones distintas del mismo «camino», dibujado de distintas formas sobre la superficie, y los caminos no equivalentes, como representaciones de «caminos» esencialmente distintos. Este convenio muestra que el conjunto de todos los caminos cerrados (omitimos ahora las comillas) que parten de un mismo punto A de la superficie es un grupo para la operación de multiplicación de caminos. El elemento unidad (neutro) del grupo es el camino cero, y el elemento inverso de un camino es el mismo camino pero recorrido en la dirección opuesta.

El grupo de caminos, en general, depende no sólo de la superficie, sino también del punto inicial A . Sin embargo, si la superficie no consta de partes separadas, es decir, si podemos unir cada dos puntos por un camino continuo sobre la superficie, entonces los grupos correspondientes a puntos distintos son isomorfos, y podemos hablar entonces de un solo grupo de caminos de la superficie S , sin especificar

el punto A . Este grupo de caminos sobre la superficie se llama también *grupo fundamental* de la superficie.

Si la superficie es un plano o una esfera, el grupo de caminos consta del elemento unidad solamente, ya que todo camino se puede contraer a un punto. Sin embargo, sobre la superficie cilíndrica circular infinita, como ya hemos visto, existen caminos cerrados que no se contraen a un punto. Puesto que sobre el cilindro todo camino cerrado que parte de A es equivalente a una potencia del camino X (Fig. 23), y las distintas potencias de X no son equivalentes, el grupo de caminos de la superficie cilíndrica es un grupo cíclico infinito. Se puede demostrar que el grupo de caminos sobre el toro (Fig. 22) consiste en caminos de la forma $U^m V^n$ ($m, n = 0, \pm 1, \pm 2, \dots$) con $UV = VU$ y $U^m V^n = U_{m_1} V_{n_1}$ sólo para $m = m_1, n = n_1$, donde recordamos que la igualdad de caminos se debe entender en el sentido de equivalencia.

La importancia del grupo de caminos se debe a la siguiente propiedad. Imaginemos, además de la superficie S , otra superficie S_1 , de forma que entre los puntos de ambas superficies se pueda establecer una correspondencia uno a uno. Una correspondencia tal es posible si, por ejemplo, la superficie S_1 se obtiene de la S por una deformación continua sin roturas ni fusiones de puntos distintos de la superficie. A cada camino sobre la superficie S corresponde un camino sobre S_1 . Además, a caminos equivalentes de S corresponden caminos equivalentes de S_1 , y al producto de caminos en S , el producto de los correspondientes en S_1 , de modo que el grupo de caminos sobre S es isomorfo al de S_1 . En otros términos, el grupo de caminos, considerado desde un punto de vista abstracto, es decir, salvo isomorfismos, es un invariante para todas las transformaciones uno a uno continuas de la superficie. Si los grupos de caminos de dos superficies son distintos, las superficies no pueden ser transformadas continuamente la una en la otra. Por ejemplo, el plano no se puede transformar sin fusiones ni roturas en la superficie cilíndrica, ya que el grupo de caminos del plano consta de un solo elemento, y el de la superficie cilíndrica es infinito.

Las propiedades de las figuras que no varían bajo transformaciones uno a uno bicontinuas se estudian en la disciplina matemática de la topología, cuyas ideas básicas fueron desarrolladas en el capítulo 17. Los invariantes de las transformaciones bicontinuas se llaman *invariantes topológicos*. El grupo de caminos es uno de los ejemplos más importantes de ellos. Es claro que el grupo de caminos se puede definir no sólo para superficies, sino también para cualquier conjunto

de puntos, siempre que se pueda hablar de caminos y de deformaciones en dichos conjuntos.

Relaciones de definición

En topología se estudian con detalle los métodos para calcular el grupo de caminos. Por lo general, resulta muy conveniente definir estos grupos por un método especial que se aplica no sólo a los grupos fundamentales en topología, sino también a la definición de grupos abstractos en general. Consiste en lo siguiente.

Sea G un grupo. Los elementos $g_1, g_2, \dots, g_n$ se llaman generadores de G si todo elemento g se puede escribir en la forma

$$g = g_{i_1}^{\alpha_1} g_{i_2}^{\alpha_2} \cdots g_{i_k}^{\alpha_k},$$

donde $i_1, i_2, \dots, i_k$ son algunos de los números $1, 2, \dots, n$; dos índices i consecutivos no pueden ser iguales; el número de factores k es arbitrario; los exponentes $\alpha_1, \alpha_2, \dots, \alpha_k$ son enteros positivos o negativos.

Para determinar el grupo G es suficiente conocer, aparte de los generadores, los productos que representan el mismo elemento del grupo y los que representan distintos elementos. Así, pues, para definir el grupo debemos determinar todas las ecuaciones de la forma

$$g_{i_1}^{\alpha_1} g_{i_2}^{\alpha_2} \cdots g_{i_k}^{\alpha_k} = g_{j_1}^{\beta_1} g_{j_2}^{\beta_2} \cdots g_{j_l}^{\beta_l}$$

que se verifican en G . Puesto que el número de tales ecuaciones es siempre infinito, sólo se dan aquellas que implican las restantes mediante los axiomas de grupo. Estas ecuaciones se llaman relaciones de definición.

Es claro que hay distintas formas de dar el mismo grupo mediante relaciones de definición.

Consideremos, por ejemplo, el grupo H con los generadores a, b y las relaciones

$$a^2 = b^3, ab = ba. \quad (10)$$

Tomando $c = ab^{-1}$, tenemos

$$a = bc, a^2 = b^2c^2, b^3 = b^2c^2, b = c^2, a = c^3.$$

Vemos, por tanto, que todos los elementos del grupo H se pueden expresar mediante el único elemento c , donde

$$a = c^3, b = c^2.$$

Puesto que la relaciones (10) se siguen inmediatamente de estas ecuaciones, no hay relaciones triviales para c . Por tanto, H es un grupo cíclico infinito cuyo elemento generador es c .

Si en un grupo podemos elegir generadores no relacionados por relaciones triviales, el grupo se llama entonces *libre*, y los generadores, generadores libres. Por ejemplo, si el grupo tiene los generadores libres a, b , entonces todo elemento suyo se puede escribir de modo único en la forma

$$a^{\alpha_0} b^{\beta_1} a^{\alpha_1} b^{\beta_2} a^{\alpha_2} \cdots b^{\beta_k} a^{\alpha_k},$$

donde $k = 0, 1, 2, \dots, n$ y los exponentes $\alpha_0, \beta_1, \alpha_1, \dots, \beta_k, \alpha_k$ son enteros positivos o negativos, salvo los «extremos» α_0 y α_k , que pueden tomar también el valor cero. Para grupos libres con un número mayor de generadores se verifican resultados similares.

Dados los generadores y las relaciones de definición de dos grupos y suponiendo que éstos no tienen elementos comunes, podemos combinar estas relaciones y obtener un nuevo grupo, que se llama producto libre de los dados.

La teoría de grupos libres, y también la más general de productos libres, ocupa un lugar importante en la teoría de grupos. Desde el punto de vista geométrico el producto libre de los grupos H_1 y H_2 es un

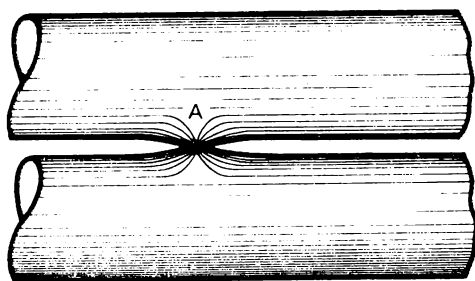


FIG. 24

grupo de caminos de la figura que se puede representar en forma de suma de dos figuras cerradas unidas en un solo punto y que tienen a H_1 y H_2 como grupos de caminos. Sabemos ya que el grupo de caminos de una superficie cilíndrica es un grupo libre con un generador. De la observación que acabamos de hacer se sigue,

por ejemplo, que el grupo de caminos de la superficie ilustrada en la figura 24 es un grupo libre con dos generadores.

De modo similar a como introdujimos el grupo fundamental de una superficie, podemos definir también el grupo fundamental de cuerpos espaciales, finitos o infinitos.

Nudos y grupos de nudos

Como ya hemos dicho, desde el punto de vista de la topología dos superficies se pueden considerar idénticas si se puede llevar la una sobre la otra mediante una transformación uno a uno bicontinua. El problema de la clasificación topológica de las superficies cerradas hace tiempo que fue resuelto. Toda superficie del espacio ordinario es topológicamente equivalente bien a la esfera, o bien a una esfera con un cierto número de asas (Fig. 25). Por ejemplo, la

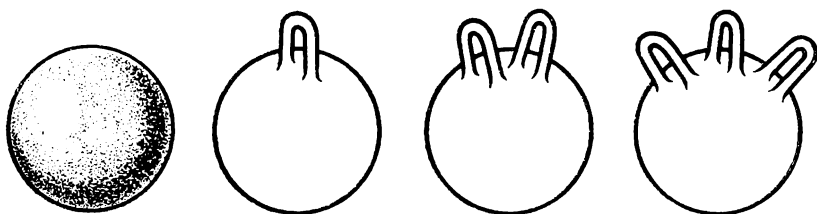


FIG. 25

superficie toroidal, ilustrada en la figura 22, se puede deformar continuamente en una esfera de una sola asa; la superficie de un cubo en la superficie de una esfera; y así sucesivamente. En vista de ello, el estudio de los grupos fundamentales de superficies cerradas no es muy interesante, ya que estas superficies están completamente clasificadas sin necesidad de dichos grupos. Sin embargo, existen problemas muy simples que no se han podido resolver en su totalidad hasta contar con los grupos fundamentales. Entre ellos figura el famoso problema de los nudos.

Un nudo es una curva cerrada del espacio ordinario tridimensional. Como muestra la figura 26, su forma puede ser muy variada. Dos nudos se dicen equivalentes si uno de ellos se puede transformar en el otro mediante un proceso continuo sin roturas ni autopenetraciones de la curva. Inmediatamente surgen dos problemas: 1) ¿Cómo podemos determinar si dos nudos dados por su proyecciones planas son o no equivalentes? 2) ¿Cómo podemos clasificar todos los nudos no equivalentes?

Ambos problemas están todavía sin resolver, pero los progresos sustanciales realizados en su solución parcial están relacionados con la teoría de grupos.

Suprimamos del espacio los puntos que pertenecen a un nudo,

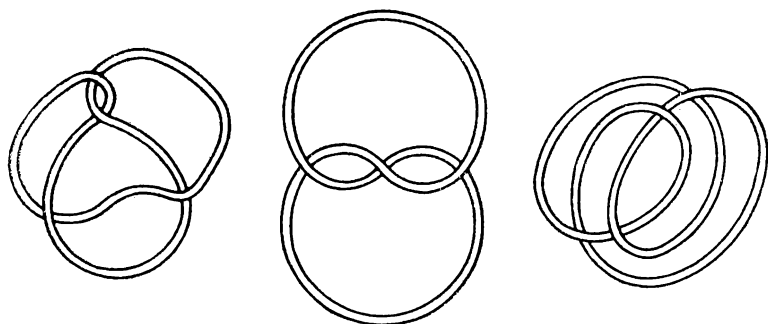


FIG. 26

y consideremos el grupo fundamental del conjunto restante de puntos. Este grupo se llama grupo del nudo. Es obvio que si los nudos son equivalentes, sus grupos son isomorfos. Por tanto, si los grupos no son isomorfos, podemos concluir que los nudos correspondientes no

son equivalentes. Por ejemplo, el grupo del nudo que se puede reducir a un círculo es un grupo cíclico. El grupo del nudo en forma de trébol es más complicado (Fig. 27). No es conmutativo y por tanto no es isomorfo al grupo de un círculo. Podemos concluir entonces que es imposible reducir un nudo de trébol a un círculo sin romperlo, hecho completamente obvio pero que requiere una demostración mediante argumentos matemáticos precisos.

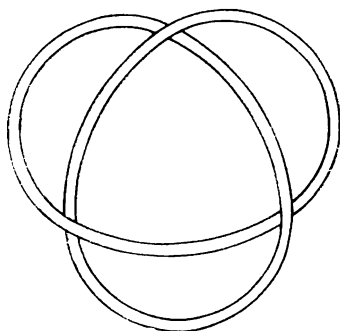


FIG. 27

Desgraciadamente, en la discusión de los grupos de nudos surgen problemas difíciles que no han sido resueltos hasta ahora. El hecho es que en topología se conocen métodos muy simples para hallar generadores y relaciones de definición para el grupo de un nudo representado en una forma dada. Mas para utilizar grupos en la comparación de distintos nudos debemos ser capaces de decidir acerca de si los grupos dados por generadores y relaciones de definición son

o no isomorfos, y la solución de este problema no se conoce hasta ahora. De hecho, el matemático soviético P. S. Novikov ha probado recientemente un notable teorema que afirma que es imposible indicar un único proceso regular (más exactamente, un algoritmo normal) que permita decidir si dos sistemas de relaciones de definición para un mismo conjunto de generadores definen o no el mismo grupo. Este teorema nos induce a dudar de la existencia de un método general uniforme para decidir sobre la equivalencia de nudos dados por sus proyecciones planas.

§9. *Representaciones y caracteres de grupos*

Los métodos de la teoría general de grupos guardan cierto parecido con los de la geometría elemental: ambas se basan en un sistema concreto de axiomas que sirve de punto de partida para la construcción de todo el contenido de la teoría. Pero el ejemplo de la geometría analítica indica hasta qué punto la aplicación de los métodos numéricos y analíticos puede mostrarse útil en el estudio de problemas geométricos.

Una aplicación de los métodos del análisis y del álgebra clásica a la teoría de grupos es la llamada teoría de representación de grupos. La geometría analítica no sólo proporciona métodos para resolver problemas geométricos por medio del análisis, sino que, a su vez, arroja luz «geométrica» sobre muchos problemas complicados del análisis. Del mismo modo, e incluso en mayor grado, la teoría de representaciones no sólo sirve de método auxiliar en el estudio de las propiedades de los grupos, sino que, al forjar un nexo entre los conceptos y problemas del análisis y la teoría de grupos, permite hallar expresiones de los resultados de la teoría de grupos en términos de relaciones numéricas y da a las relaciones analíticas una interpretación mediante grupos. Gran parte de las aplicaciones actuales de la teoría de grupos a la física está relacionada precisamente con la teoría de representaciones.

Representaciones de grupos por matrices

En álgebra lineal (ver capítulo 15) hemos discutido la operación de multiplicación de matrices. Esta operación es asociativa, pero en general no conmutativa. Las matrices no singulares cuadradas de un cierto orden forman un grupo para la multiplicación; ya que el producto de dos matrices no singulares es no singular, existe un ele-

mento unidad que es la matriz unidad, y para toda matriz no singular existe una matriz inversa también no singular.

Supongamos dado un cierto grupo G y que a cada uno de sus elementos g se le asocia una matriz de números complejos A_g de orden n tal que si multiplicamos dos elementos del grupo, sus matrices correspondientes también se multiplican: $A_{gh} = A_g \cdot A_h$. Decimos entonces que tenemos una representación del grupo G mediante matrices de grado n . Por lo general, la frase «mediante matrices» se omite y se habla simplemente de una representación de grado n de G . Una representación de grado n de un grupo G no es más que un homomorfismo del grupo G en el grupo de las matrices no singulares de grado n . De las propiedades generales de los homomorfismos se sigue que en toda representación el elemento neutro de G se aplica en la matriz unidad, y los elementos inversos, en matrices inversas.

Las matrices de orden 1 son simplemente números complejos. Por tanto, una representación de grado uno de G es una correspondencia que asigna a cada elemento de G un número complejo, y al producto de dos elementos de G , el producto de los números complejos correspondientes. Por ejemplo, la aplicación que a cada permutación par asocia el número 1 y a cada permutación impar el número -1 es una representación de grado 1.

Asociando a todo elemento de un grupo G la matriz unidad E de grado n , obtenemos una representación de G que recibe el nombre de representación unidad de grado n . Si G es un grupo finito que contiene más de un elemento, tiene también necesariamente un número infinito de representaciones, además de la representación unidad de distintos grados. Los métodos para hallarlas se indican a continuación.

Si conocemos una representación del grupo G , podemos obtener un conjunto infinito de otras representaciones. Para ello, sea $g \rightarrow A_g$ la representación dada de G mediante matrices de grado n . Tomamos una matriz no singular arbitraria P del mismo grado n y ponemos $B_g = P^{-1}A_gP$. La correspondiente $g \rightarrow B_g$ es otra representación de G , ya que

$$B_{gh} = P^{-1}A_{gh}P = P^{-1}A_gA_hP = P^{-1}A_gPP^{-1}A_hP = B_gB_h.$$

Las representaciones obtenidas así, a partir de una representación dada, mediante la elección de distintas matrices P , se dice que son *equivalentes* a la dada. En la teoría de representaciones, las represen-

taciones equivalentes no se consideran como esencialmente distintas.

Otro método de encontrar nuevas representaciones es la suma directa, que consiste en lo siguiente: sean $g \rightarrow A_g$, $g \rightarrow B_g$ representaciones arbitrarias de un grupo G de grados m y n respectivamente. Consideremos la aplicación

$$g \rightarrow \begin{bmatrix} A_g & 0 \\ 0 & B_g \end{bmatrix}.$$

Por la regla de multiplicación de matrices (ver capítulo 15) obtenemos

$$gh \rightarrow \begin{bmatrix} A_{gh} & 0 \\ 0 & B_{gh} \end{bmatrix} = \begin{bmatrix} A_g A_h & 0 \\ 0 & B_g B_h \end{bmatrix} = \begin{bmatrix} A_g & 0 \\ 0 & B_g \end{bmatrix} \begin{bmatrix} A_h & 0 \\ 0 & B_h \end{bmatrix};$$

es decir, la aplicación es otra representación de G . Se llama suma de las dos representaciones dadas y se denota por $A_g + B_g$. Transponiendo los sumandos se obtiene otra representación

$$g \rightarrow \begin{bmatrix} B_g & 0 \\ 0 & A_g \end{bmatrix},$$

que es, no obstante, equivalente a la anterior. Por tanto, si no distinguimos entre representaciones equivalentes, la suma de representaciones es una operación conmutativa. Es fácil comprobar que, con la misma condición, la suma es también una operación asociativa. A partir de un cierto número de representaciones $A_g, B_g, C_g, \dots$ de un grupo G podemos obtener, mediante la suma, representaciones de grados cada vez mayores: $A_g + B_g + C_g$, $A_g + A_g + A_g + A_g$, y así sucesivamente.

Por ejemplo, los números $1, -1, i, -i$ forman grupo para la multiplicación. Asociando cada número de este grupo consigo mismo, obtenemos una representación de grado 1. Tomemos una segunda representación $1 \rightarrow 1, -1 \rightarrow -1, i \rightarrow -i, -i \rightarrow i$. La suma de estas representaciones es la aplicación

$$1 \rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; -1 \rightarrow \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}; i \rightarrow \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}; -i \rightarrow \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix}.$$

Transformándolas mediante la matriz

$$P = \begin{bmatrix} 1 & i \\ 1 & -i \end{bmatrix}$$

obtenemos la representación equivalente

$$1 \rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, -1 \rightarrow \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, i \rightarrow \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, -i \rightarrow \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Es interesante observar que todas las matrices de esta representación son reales. Supongamos ahora que todas las matrices de una representación de grado n de un grupo G tienen la forma

$$g \rightarrow A_g = \begin{bmatrix} B_g & C_g \\ 0 & D_g \end{bmatrix},$$

donde B_g, C_g, D_g son matrices cuadradas y el rectángulo inferior izquierdo de A_g está formado por ceros. Multiplicando las matrices A_g y A_h obtenemos

$$A_{gh} = A_g A_h = \begin{bmatrix} B_g B_h & B_g C_h + C_g D_h \\ 0 & D_g D_h \end{bmatrix};$$

esto es, $B_{gh} = B_g B_h$, $D_{gh} = D_g D_h$. Esto prueba que las aplicaciones $g \rightarrow B_g$ y $g \rightarrow D_g$ son también representaciones de G , pero de menor grado. A_g se llama una representación graduada de G , y toda representación equivalente a ella se dice *reducible*. Una representación que no es equivalente a ninguna representación graduada se llama *irreducible*.

Si en todas las matrices de A_g no sólo el rectángulo inferior de la izquierda, sino también el superior de la derecha son iguales a cero, se dice que A_g se divide en la suma de las dos representaciones B_g, D_g . Una representación equivalente a una suma de representaciones irreducibles se llama *completamente reducible*.

En la teoría de grupos se prueba que toda representación de un grupo finito es completamente reducible³. Se sigue de aquí que para

³ Recordamos que estamos considerando representaciones de grupos mediante matrices de números complejos.

hallar todas las representaciones de un grupo finito es suficiente conocer sus representaciones irreducibles, ya que todas las demás son equivalentes a sumas de irreducibles.

El cálculo práctico de las representaciones irreducibles de un grupo finito arbitrario es, en general, un trabajo fracamente complicado que sólo se puede resolver explícitamente para algunas clases de grupos finitos; por ejemplo, para grupos conmutativos, grupos simétricos y algunos otros, se han estudiado con detalle, aunque desde un punto de vista teórico, las propiedades de las representaciones de los grupos finitos.

Todo grupo finito tiene una representación «regular» que se construye como sigue. Supongamos que $g_1, g_2, \dots, g_n$ son los elementos del grupo dado G , enumerados en un orden cualquiera, y que

$$g_i g_k = g_{i_k} \quad (i, k = 1, 2, \dots, n).$$

Tomando un valor fijo para k , formamos la matriz de grado n que tiene un 1 en el lugar i y ceros en los lugares restantes ($i = 1, 2, \dots, n$) y la denotamos por R_{g_k} . La correspondencia $g \rightarrow R_{g_k}$ ($k = 1, 2, \dots, n$) se llama *representación regular* de G . Un simple cálculo basta para demostrar que es una representación.

Se puede probar también que cambiando la numeración de los elementos del grupo llegamos a una representación equivalente y que, en consecuencia, salvo una equivalencia todo grupo finito tiene una sola representación regular.

Formulemos brevemente los teoremas de la teoría de representación de grupos finitos. El número de representaciones irreducibles distintas (no equivalentes) de un grupo finito es finito e igual al número de clases de elementos conjugados del grupo (ver §6). El grado de una representación irreducible es necesariamente un divisor del orden del grupo, y la representación regular es equivalente a una suma de todas las representaciones irreducibles no equivalentes en las que todo sumando irreducible se repite tantas veces como indica su grado.

Esto implica la interesante relación siguiente entre el orden de un grupo finito y los grados de sus representaciones irreducibles.

Denotamos por n el número de elementos del grupo G ; el número de clases de elementos conjugados por, k , y los grados de sus representaciones irreducibles, por $n_1, n_2, \dots, n_k$, respectivamente. Es evidente (por la construcción de la representación regular) que la representación regular tiene de grado n . Además, puesto que la representación regular es equivalente a una suma de n_1 representaciones

equivalentes a la primera representación irreducible, más n_2 equivalentes a la segunda, y así sucesivamente, y puesto que al sumar representaciones sus grados se suman, tenemos la siguiente ecuación:

$$n = n_1^2 + n_2^2 + \cdots + n_k^2. \quad (11)$$

Asociando a todo elemento del grupo el número 1 obtenemos las representación irreducible trivial de grado 1, que todo grupo posee. Si en la fórmula (11) tomamos n_1 como el grado de esta representación unidad, podemos escribir de nuevo (11) en la forma equivalente:

$$n = 1 + n_2^2 + \cdots + n_k^2,$$

donde $n_2, \dots, n_k$ denotan ahora los grados de las representaciones irreducibles no triviales.

Teniendo en cuenta que $n_2, \dots, n_k$ deben ser divisores de n , podemos, en ocasiones (si conocemos k), hallar $n_2, \dots, n_k$ a partir de la ecuación (11) solamente. Por ejemplo, el grupo simétrico S_3 de las permutaciones de tres elementos tiene tres clases de permutaciones conjugadas: (1); (1,2), (1,3), (2,3); (1,2,3), (1,3,2). Para $n = 6$, $k = 3$ la ecuación (11) admite sólo un sistema de soluciones: $6 = 1^2 + 1^2 + 2^2$. Por tanto, S_3 tiene dos representaciones distintas de grado 1 y una representación irreducible de grado 2.

Otro ejemplo son los grupos abelianos finitos. En este caso cada elemento forma una clase individual. Por tanto $k = n$, y de la fórmula (11) se sigue que $n_1 = n_2 = \cdots = n_k = 1$, esto es, todas las representaciones irreducibles de estos grupos son de grado 1 y su número es igual al orden del grupo.

Las representaciones irreducibles de los grupos abelianos se llaman también caracteres, mientras que para toda representación de un grupo no abeliano la palabra «carácter» se utiliza para designar el conjunto de las llamadas trazas (la suma de los elementos diagonales) de las matrices que forman la representación. Los caracteres de grupos finitos tienen importantes propiedades y relaciones. La investigación de las representaciones y caracteres de grupos ha enriquecido la teoría de grupos con interesantes resultados generales, que han encontrado una extensa aplicación en la física teórica contemporánea.

§10. *La teoría general de grupos*

Hemos mencionado ya que casi a lo largo de todo el siglo pasado la teoría de grupos se desarrolló principalmente como una teoría de grupos de transformaciones. Sin embargo, fue haciéndose cada vez más claro que el estudio de los grupos como tales es fundamental y que el estudio de los grupos de transformaciones se puede reducir al de los grupos abstractos y sus subgrupos. El paso de la teoría de grupos de transformaciones a la de grupos abstractos se efectuó primeramente en la teoría de grupos finitos, pero el rápido desarrollo de la teoría de grupos de Lie y la irrupción de la teoría de grupos en la topología hizo necesario crear la teoría general de grupos, en la que los grupos finitos son considerados solamente como un caso especial.

El primer texto sobre la teoría de grupos en el que se adoptó este punto de vista fue el libro de O. Ju. Smidt, publicado en Kiev en 1916. Smidt obtuvo también durante los años veinte un importante teorema sobre grupos infinitos que se convirtió en el punto de partida en las investigaciones de otros muchos algebristas soviéticos. Gracias a los trabajos de O. Ju. Smidt y P. S. Aleksandrov, que contribuyeron no poco a la divulgación de las ideas del álgebra contemporánea, se formó en Moscú una gran escuela de teoría de grupos que más tarde quedó bajo la dirección de su discípulo A. G. Kuros. Kuros se dio a conocer, en particular, por la demostración de un teorema que afirma que todo subgrupo de un producto libre es también un producto libre de subgrupos isomorfos a subgrupos apropiados de los factores y, posiblemente, un subgrupo libre distinto. Más tarde publicó una monografía sobre la teoría de grupos en la cual se da por primera vez una exposición sistemática del rico material de resultados obtenidos en la teoría general de grupos. Esta monografía es todavía el libro de texto más completo de toda la literatura sobre teoría general de grupos y se ha hecho internacionalmente famosa.

Siguiendo la pauta de la escuela de Moscú, los algebristas de Ленинград y otras ciudades se interesaron en la teoría general de grupos, colaborando con sus propias contribuciones a su desarrollo. Las investigaciones llevadas a cabo hasta el presente en la Unión Soviética abarcan todas las ramas esenciales de la teoría de grupos, y los resultados obtenidos por los matemáticos soviéticos han ejercido más de una vez una influencia decisiva en el desarrollo de esta teoría.

§11. *Números hipercomplejos*

En la resolución de problemas prácticos por métodos algebraicos se suele llegar, en los casos más simples, a una o varias ecuaciones, a partir de las cuales debemos hallar los valores de las incógnitas. Las incógnitas, en este contexto, son características cuantitativas de los objetos que se estudian; las ecuaciones se construyen mediante un análisis de las relaciones reales que existen entre los objetos.

Esta es la situación en los casos en que se trata de las cantidades más simples —tales como masa, volumen o distancia— que se pueden caracterizar por un solo número. Sin embargo, en muchos problemas concretos, encontramos objetos que no se pueden caracterizar por un solo número. Es más, en el desarrollo de la tecnología las entidades más importantes son de una naturaleza muy complicada y requieren para su caracterización varios números, e incluso infinitos. Cantidades físicas tan importantes como las fuerzas, velocidades o aceleraciones se caracterizan mediante segmentos dirigidos y requieren tres números. Es bien sabido que la posición de un punto en el espacio se caracteriza por tres números, la posición de un plano también por tres, la de una recta por cuatro, y la de un sólido rígido por seis números. Por tanto, cuando deseamos resolver por métodos algebraicos problemas referentes a objetos más complicados, obtenemos ecuaciones con un gran número de incógnitas, por lo que a menudo resulta mucho más pesado analizarlas que resolver el problema haciendo uso de sus propiedades geométricas y físicas. De aquí la idea de tratar de caracterizar los objetos de naturaleza complicada no por sistemas de números ordinarios, sino por números más generales y complicados con los que se puedan realizar operaciones similares a las aritméticas. Este planteamiento del problema era el más natural, ya que la historia de la ciencia ha mostrado, no la invariabilidad del concepto de número, sino su flexibilidad; el gradual enriquecimiento del campo de los números, desde los números naturales hasta los números racionales, luego los algebraicos, después los reales (rationales o irracionales) y finalmente los números complejos.

Números complejos

El lector está ya familiarizado desde el capítulo 4 con las propiedades de los números complejos y sus aplicaciones más sencillas. Ahora nos interesa solamente la fundamentación del concepto de

número complejo. Cuando estudiamos los números reales ordinarios, observamos que la raíz cuadrada de un número negativo no tiene sentido, ya que el cuadrado de todo número real es positivo o cero. Entonces se procede a mostrar cómo las necesidades de la ciencia obligaron a los matemáticos a considerar expresiones de la forma $a + b\sqrt{-1}$ como una clase especial de número que se empezó a llamar imaginario por oposición a los números reales ordinarios. Si se supone que estos números imaginarios están sujetos a las mismas leyes de las operaciones aritméticas que los números reales ordinarios, entonces todas las raíces cuadradas de números negativos se pueden expresar en función de la cantidad $i = \sqrt{-1}$, y el resultado de las operaciones aritméticas realizadas un número finito de veces con números reales o imaginarios se puede representar siempre en la forma $a + bi$, donde a y b son números reales.

Evidentemente, esta definición de número imaginario va contra el sentido común en alto grado: primero se afirma que las expresiones $\sqrt{-1}$, $\sqrt{-2}$, etc., no tienen sentido, y luego se propone que estas expresiones sin sentido sean llamadas números imaginarios. Esta circunstancia hizo dudar a los matemáticos de los siglos XVII y XVIII de la validez del uso de los números complejos. Sin embargo, estas dudas se disiparon a comienzos del siglo XIX cuando se encontró una representación de los números complejos mediante puntos del plano. Algo más tarde, el matemático húngaro Bolyai y el irlandés Hamilton dieron una fundamentación puramente aritmética a la teoría de los números complejos. El procedimiento es como sigue.

En lugar de hablar de números $a + bi$, lo haremos solamente de pares de números reales (a, b) . Dos pares se consideran iguales si sus primeros y segundos términos son iguales, es decir, $(a, b) = (c, d)$ si, y sólo si, $a = c$ y $b = d$. La suma y multiplicación de pares se definen por las fórmulas

$$(a, b) + (c, d) = (a + c, b + d); (a, b) \cdot (c, d) = (ac - bd, ad + bc).$$

Tenemos, por ejemplo,

$$\begin{aligned} (2, 3) + (1, -2) &= (3, 1) & (2, 3)(1, -2) &= (8, -1), \\ (3, 0) + (2, 0) &= (5, 0). & (3, 0)(2, 0) &= (6, 0). \end{aligned}$$

Estos ejemplos muestran, en particular, que las operaciones aritméticas con parejas cuyas primeras componentes son cero se reducen

a las operaciones con los primeros términos únicamente, de modo que estas parejas se pueden denotar por sus primeros números solamente. Si introducimos la notación i para el par $(0, 1)$ tenemos entonces

$$(a, b) = a(1, 0) + b(0, 1) = a + bi,$$

$$i^2 = (0, 1)(0, 1) = (-1, 0) = -1;$$

es decir, aparecen las notaciones usuales para números complejos.

Así, pues, desde este punto de vista los números complejos son pares ordinarios de números reales y las operaciones con ellos son solamente una clase especial de operaciones con pares de números reales.

Números hipercomplejos

Diversas aplicaciones felices de los números complejos indujeron a los matemáticos, ya en las primeras décadas del siglo XIX, a centrar la atención sobre el problema de si se podían o no construir números complejos superiores representados por ternas, cuaternas, etc., de números reales, del mismo modo que los números complejos ordinarios tenían la forma de pares de números reales. Desde mediados del siglo pasado en adelante se han estudiado muchos sistemas especiales de tales números complejos superiores o números hipercomplejos, y a finales del pasado siglo y comienzos del actual se desarrolló una teoría general de números hipercomplejos que ha encontrado importantes aplicaciones en campos colindantes de las matemáticas y la física.

Recibe el nombre de número hipercomplejo de rango n un número que se puede representar por una colección de n números reales $(a_1, a_2, \dots, a_n)$, que por el momento llamaremos coordenadas del número. Los números hipercomplejos $(a_1, a_2, \dots, a_n)$ y $(b_1, b_2, \dots, b_n)$ se dicen iguales si sus correspondientes coordenadas son iguales, esto es, si $a_1 = b_1, a_2 = b_2, \dots, a_n = b_n$. Definimos la operación de suma por la fórmula

$$(a_1, a_2, \dots, a_n) + (b_1, b_2, \dots, b_n) = (a_1 + b_1, a_2 + b_2, \dots, a_n + b_n),$$

análoga a la de la suma de números complejos.

También es natural introducir la operación de multiplicación de un número hipercomplejo por un número real: por definición ponemos

$$a(a_1, a_2, \dots, a_n) = (aa_1, aa_2, \dots, aa_n).$$

Tenemos que definir ahora la operación de multiplicación de dos números hipercomplejos de forma que el resultado sea otro número hipercomplejo.

Extender la definición de multiplicación de números complejos ordinarios al caso general es laborioso. Se puede hacer de varias formas, obteniendo así distintos sistemas de números hipercomplejos. Debemos, por tanto, aclarar cuál de esas definiciones se va a tomar.

Indudablemente, sería conveniente que las operaciones con números hipercomplejos que vamos a definir tuvieran las mismas propiedades que las operaciones ordinarias con números reales. ¿Pero cuáles son esas propiedades de las operaciones ordinarias?

En una cuidadosa discusión de las propiedades de los números y de las operaciones con ellos que se utilizan más frecuentemente en álgebra, es fácil observar que aquéllas se reducen a las siguientes:

1. La suma de dos números está unívocamente determinada.

2. Para cualesquiera dos números, su producto está determinado de manera única.

3. Existe un número cero con la propiedad $a + 0 = a$ para todo a .

4. Para todo número a existe el número opuesto x que satisface la ecuación $a + x = 0$.

5. La adición es conmutativa

$$a + b = b + a.$$

6. La adición tiene la propiedad asociativa

$$(a + b) + c = a + (b + c).$$

7. La multiplicación es conmutativa

$$ab = ba.$$

8. La multiplicación es asociativa

$$(ab) \cdot c = a \cdot (bc).$$

9. La multiplicación es distributiva respecto a la suma

$$a(b + c) = ab + ac, \quad (b + c)a = ba + ca.$$

10. Para todo a y todo $b \neq 0$, existe un único número x que verifica la ecuación $bx = a$.

Las propiedades 1-10 han sido seleccionadas como resultado de un cuidadoso análisis; el desarrollo de las matemáticas en el siglo pasado demostró su gran importancia. Hoy día, todo sistema de cantidades que satisface las propiedades 1-10 recibe el nombre de *cuerpo*. Ejemplos de cuerpos son: el conjunto de los números racionales, el conjunto de los números reales y el conjunto de los números complejos; en cada uno de estos casos, los elementos del conjunto se pueden sumar y multiplicar, el resultado es un número del mismo conjunto y se verifican las propiedades 1-10. Aparte de estos tres cuerpos muy importantes se pueden determinar una infinidad de otros cuerpos formados por números. Pero, además de los cuerpos formados por números, tienen también mucho interés los constituidos por objetos de otra naturaleza. Por ejemplo, en la escuela aprendemos a operar con las llamadas funciones algebraicas, es decir, fracciones en las que el numerador y el denominador son polinomios en ciertas letras. Las fracciones algebraicas se pueden sumar, multiplicar y dividir, y estas operaciones tienen las propiedades 1-10. Por consiguiente, las fracciones algebraicas forman un sistema de objetos que es un cuerpo. Podríamos dar otros muchos ejemplos de cuerpos formados por objetos de naturaleza más complicada. En vista de la importancia de las propiedades 1-10 que definen un cuerpo, la formulación original del problema fue encontrar una operación sobre los números hipercomplejos tal que éstos formaran un cuerpo. En caso de éxito se podría tratar de obtener nuevos números complejos más generales aún. Sin embargo, ya a comienzos del siglo pasado se descubrió que esto es sólo posible para números hipercomplejos de rango 2 y que sólo los números complejos ordinarios podrían obtenerse de esta forma. Este resultado demostró que los números complejos ocupan una posición muy especial y que es imposible obtener una extensión del sistema de números más allá de ellos, en el caso, insistentes, de exigir las propiedades 1-10.

Por tanto, en cualquier intento posterior de construir sistemas superiores de números fue necesario omitir una o varias de las propiedades 1-10.

Cuaternos

Históricamente, el primer sistema hipercomplejo que se discutió en matemáticas fue el sistema de cuaternos, esto es, «números cuádruples», que fueron introducidos por el matemático irlandés Hamilton a mediados del siglo pasado. Este sistema satisface todas las pro-

iedades de la 1 a la 10 excepto la 7 (conmutatividad de la multiplicación).

Los cuaternos se pueden describir como sigue. Para los entes $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, $(0, 0, 0, 1)$ introducimos las abreviaturas $1, i, j, k$. Entonces por la ecuación

$$(a, b, c, d) = a(1, 0, 0, 0) + b(0, 1, 0, 0) + c(0, 0, 1, 0) + d(0, 0, 0, 1)$$

todo cuaterno se puede escribir unívocamente en la forma

$$(a, b, c, d) = a \cdot 1 + b \cdot i + c \cdot j + d \cdot k.$$

El cuaterno 1 juega el papel de elemento unidad en el sistema que se va a construir; esto es, suponemos que $1 \cdot a = a \cdot 1 = a$ para todo cuaterno a . Además, por definición escribimos: $i^2 = j^2 = k^2 = -1$;

$$ij = -ji = k,$$

$$ik = -ki = -j,$$

$$jk = -kj = i.$$

Esta tabla de multiplicación resulta fácil de memorizar mediante la figura 28, en la que los puntos i, j, k sobre el círculo representan los correspondientes cuaternos i, j, k . El producto de dos cuaternos adyacentes es igual al tercero si para ir del primero al segundo en la figura, por el camino más corto, se hace en el mismo sentido de las agujas del reloj. Conocida la tabla de multiplicar de los cuaternos i, j, k , se puede efectuar el producto de cuaternos arbitrarios utilizando la ley distributiva 9. En efecto:

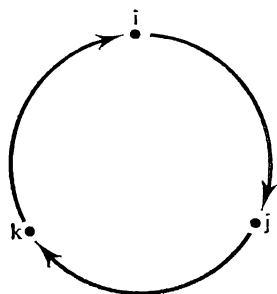


FIG. 28

$$\begin{aligned} (a \cdot 1 + b \cdot i + c \cdot j + d \cdot k)(a_1 \cdot 1 + b_1 \cdot i + c_1 \cdot j + d_1 \cdot k) = \\ = aa_1 \cdot 1 + ab_1 \cdot i + ac_1 \cdot j + ad_1 \cdot k + \\ + ba_1 \cdot i + bb_1 \cdot ii + bc_1 \cdot ij + bd_1 \cdot ik + \\ + ca_1 \cdot j + cb_1 \cdot ji + cc_1 \cdot jj + cd_1 \cdot jk + \\ + da_1 \cdot k + db_1 \cdot ki + dc_1 \cdot kj + dd_1 \cdot kk = \\ = (aa_1 - bb_1 - cc_1 - dd_1) \cdot 1 + (ab_1 + ba_1 + cd_1 - dc_1) \cdot i + \\ + (ac_1 + ca_1 - bd_1 + db_1) \cdot j + (ad_1 + da_1 + bc_1 - cb_1) \cdot k. \end{aligned}$$

El factor 1 en el primer término de un cuatérno se suele omitir y en lugar de $a \cdot 1$ se escribe a . Las ecuaciones $ij = -ji$, $jk = -kj$, $ik = -ki$ muestran que la multiplicación no es conmutativa. En este caso el multiplicando y el multiplicador no gozan de igual *estatus*. Por tanto, en el cálculo con cuatérnos hay que tener muy en cuenta el orden de los factores. Por otra parte, las operaciones con cuatérnos no presentan dificultad alguna. En particular, se cumple la propiedad asociativa para la multiplicación. Se puede comprobar fácilmente para los cuatérnos de la base i, j, k por medio de la tabla de multiplicar; el paso al caso general es inmediato.

El número a del cuatérno $a + bi + cj + dk$ se dice que es su parte real o escalar, y la suma $bi + cj + dk$ es la parte vectorial. Los cuatérnos $a + bi + cj + dk$ y $a - bi - cj - dk$, que difieren sólo en el signo de la parte vectorial, se dice que son conjugados. Evidentemente, la suma de dos cuatérnos conjugados es un número real. Además, al multiplicar cuatérnos conjugados por la fórmula anterior obtenemos

$$(a + bi + cj + dk) = (a - bi - cj - dk) = a^2 + b^2 + c^2 + d^2; \quad (12)$$

es decir, el producto de cuatérnos conjugados es también un número real.

La suma de los cuadrados de las coeficientes $a^2 + b^2 + c^2 + d^2$ de un cuatérno se llama norma del cuatérno. Puesto que el cuadrado de todo número real es no negativo, la norma de todo cuatérno es también no negativa, y es igual a cero sólo para el cuatérno nulo.

La fórmula (12) muestra que el producto de todo cuatérno por su conjugado es igual a su norma.

Denotaremos con un asterisco el cuatérno conjugado de uno dado. Entonces el producto verifica la siguiente fórmula:

$$(a\beta)^* = \beta^* a^*.$$

Tenemos así una interesante consecuencia: la norma del producto de cuatérnos es igual al producto de las normas de los factores. Pues por la fórmula precedente tenemos

$$\text{norm}(a\beta) = (a\beta)(a\beta)^* = a\beta\beta^*a^* = (aa^*)(\beta\beta^*) = \text{norm } a \cdot \text{norm } \beta.$$

Las propiedades de la norma permiten dar una fórmula muy

sencilla para la división de cuaternos. Sea $\alpha = a + bi + cj + dk$ un cuaterno arbitrario distinto de cero. Entonces

$$\begin{aligned} (a + bi + cj + dk) \frac{1}{a^2 + b^2 + c^2 + d^2} (a - bi - cj - dk) &= \\ = \frac{1}{a^2 + b^2 + c^2 + d^2} (a^2 + b^2 + c^2 + d^2) &= 1; \end{aligned}$$

es decir, el cuaterno

$$\frac{1}{a^2 + b^2 + c^2 + d^2} (a - bi - cj - dk) = \alpha^{-1}$$

es el inverso del cuaterno dado α .

Sabiendo cómo hallar el inverso de un cuaterno es fácil hallar ahora el cociente de dos cuaternos. Supongamos dados dos cuaternos α , β y que el primero de ellos es distinto de cero. Entonces los cocientes obtenidos al dividir β por α deben ser las soluciones de las ecuaciones

$$\alpha x = \beta, \quad y \alpha = \beta.$$

Multiplicando a la izquierda ambos miembros de la primera ecuación por el cuaterno inverso α^{-1} , obtenemos

$$x = \alpha^{-1} \beta.$$

Multiplicando a la izquierda ambos miembros de la segunda por α^{-1} obtenemos

$$y = \beta \alpha^{-1}.$$

Puesto que los productos $\alpha^{-1} \beta$ y $\beta \alpha^{-1}$ son en general distintos, debemos distinguir entre dos divisiones, a la derecha y a la izquierda; ambas son siempre posibles, excepto cuando el divisor es cero.

El álgebra de vectores

Aunque las operaciones con cuaternos son en muchos aspectos análogas a las de números complejos, la ausencia de la ley conmutativa en la multiplicación hace que las propiedades de los cuaternos sean muy diferentes a las de los números complejos. Por ejemplo,

sabemos por el álgebra de números complejos que una ecuación cuadrática tiene dos raíces. Pero si consideramos la ecuación

$$x^2 + 1 = 0$$

en el dominio de los cuaternos, podemos encontrar fácilmente 6 raíces: $\pm i, \pm j, \pm k$, y un análisis más preciso demuestra que existe incluso un número infinito de soluciones. Esta circunstancia prohíbe definitivamente el uso de los cuaternos en matemáticas, y a pesar de los numerosos intentos de Hamilton y otros matemáticos para introducirlos en distintas ramas de matemáticas y física, el papel de los cuaternos es hoy día muy modesto, no pudiéndose comparar al de los números complejos.

Sin embargo, los cuaternos han impulsado el desarrollo del álgebra de vectores, que es un instrumento indispensable en la tecnología moderna y en física. El hecho es que en mecánica y física los conceptos de velocidad, aceleración, fuerza y demás, que requieren tres números para su caracterización, juegan un papel esencial. Hemos visto anteriormente que todo cuaterno se puede considerar como suma de un número real a y una parte vectorial $bi + cj + dk$. Puesto que la parte vectorial de un cuaterno está determinada por tres números, las cantidades físicas más importantes se pueden caracterizar mediante partes vectoriales de cuaternos.

Geométricamente, la parte vectorial $bi + cj + dk$ de un cuaterno $a + bi + cj + dk$ puede considerarse como un vector que une el origen de coordenadas de un sistema cartesiano de coordenadas rectangulares con el punto cuyas proyecciones sobre estos ejes son los números b, c, d , respectivamente. En consecuencia, todo cuaterno se puede representar geométricamente como un agregado de un número y un vector en el espacio. Veamos cómo se deben interpretar las operaciones con cuaternos.

Tomamos dos cuaternos vectoriales $xi + yj + zk$ y $x_1i + y_1j + z_1k$ cuyas partes escalares son iguales a cero. Geométricamente vienen ilustrados por vectores con origen en el origen de coordenadas. La suma de estos cuaternos es otro cuaterno vectorial $(x + x_1)i + (y + y_1)j + (z + z_1)k$. Es fácil ver que el vector que representa esta suma es la diagonal del paralelogramo construido sobre los dos vectores. Así, pues, la suma de cuaternos vectoriales corresponde a la suma de vectores por la conocida ley del paralelogramo. Análogamente, si multiplicamos un cuaterno vectorial por un número real

arbitrario, el vector representante queda también multiplicado por el mismo número.

Cuando multiplicamos cuaternos llegamos a una situación diferente. En efecto,

$$(xi + yj + zk)(x_1i + y_1j + z_1k) = \\ = -xx_1 - yy_1 - zz_1 + (yz_1 - y_1z)i + (zx_1 - z_1x)j + (xy_1 - x_1y)k;$$

es decir, si multiplicamos dos cuaternos vectoriales obtenemos un cuaterno completo con una parte escalar y otra vectorial.

La parte escalar del producto de cuaternos vectoriales tomada con signo opuesto se llama producto escalar de los vectores que representan los cuaternos dados, y el vector que representa la parte vectorial del producto es el producto vectorial de los cuaternos dados. El producto escalar de los vectores α y β se denota habitualmente por $(\alpha\beta)$ o simplemente por $\alpha\beta$, y el producto vectorial de los mismos vectores por $[\alpha\beta]$. Sean i, j, k los vectores correspondientes a los cuaternos i, j, k , esto es, vectores de longitud unidad a lo largo de los ejes coordenados. Por definición, si $\alpha = xi + yj + zk$, y $\beta = x_1i + y_1j + z_1k$, entonces

$$(\alpha\beta) = xx_1 + yy_1 + zz_1, \\ [\alpha\beta] = (yz_1 - y_1z)i + (zx_1 - z_1x)j + (xy_1 - x_1y)k.$$

Mediante estas últimas fórmulas es fácil dar una interpretación geométrica del producto escalar y vectorial de vectores. Resulta que el producto escalar de dos vectores es igual al producto de sus longitudes por el coseno del ángulo que forman; el producto vectorial de dos vectores es un vector de longitud igual al área del paralelogramo construido sobre ellos, de dirección perpendicular al plano por los vectores de sentido hacia el lado desde donde la rotación del primer vector al segundo se ve igual que la rotación del eje x al eje y vista desde el eje z .

Hoy en día, en mecánica y en física no se usan, en general, operaciones con cuaternos, considerándose solamente operaciones con vectores, que se definen de una forma puramente geométrica de acuerdo con los resultados citados.

Como conclusión deseamos señalar un problema de mecánica que se resuelve por cuaternos de una forma particularmente elegante.

Su solución fue, de hecho, uno de los motivos fundamentales para el descubrimiento de los cuaternos.

Imaginemos un sólido rígido que se le hace girar primero un ángulo ϕ en cierto sentido alrededor de un cierto eje OA que pasa por el punto O , y después otro ángulo ϕ_1 alrededor de otro eje OB que pasa por el mismo punto. La cuestión es: ¿alrededor de qué eje y con qué ángulo debe girar el cuerpo para pasar con un solo movimiento de la primera a la tercera posición? Este es un problema muy conocido en mecánica: el problema de la adición de rotaciones finitas. Ciertamente, se puede resolver por métodos de geometría analítica ordinaria, como lo hizo Euler ya en el siglo XVIII. Sin embargo, la solución toma una forma más elegante mediante cuaternos.

Sean $\xi = xi + yj + zk$ y $a = a + bi + cj + dk$ dos cuaternos, el primero de los cuales se considera variable y el segundo fijo. La expresión $a^{-1}\xi a$ es un cuatérno vectorial, como se puede verificar fácilmente por un simple cálculo. Si los cuaternos ξ , $a^{-1}\xi a$ y la parte vectorial de a se representan ahora por los vectores $\vec{\xi}$, $\vec{\xi}_1$, $\vec{a}$, resulta entonces que el vector $\vec{\xi}_1$ se obtiene geoméricamente de $\vec{\xi}$ por una rotación alrededor de un eje que contiene al vector $\vec{a}$ y de ángulo ϕ definido por la fórmula

$$\cos \frac{\phi}{2} = \frac{a}{\sqrt{a^2 + b^2 + c^2 + d^2}}.$$

Podemos decir, por tanto, que el cuatérno $a = a + bi + cj + dk$ representa la rotación de ángulo ϕ alrededor del eje $\vec{a} = bi + cj + dk$.

Inversamente, conocido un eje de rotación y un ángulo ϕ se puede encontrar el cuatérno que la representa. Existe un conjunto infinito de cuaternos, pero difieren entre sí solamente en un factor numérico.

Consideremos ahora otra rotación de ángulo ϕ_1 alrededor de un eje $\vec{\beta} = b_1i + c_1j + d_1k$. Representemos esta rotación por el cuatérno $\beta = a_1 + b_1i + c_1j + d_1k$. La primera rotación transforma un vector arbitrario $\xi = xi + yj + zk$ en el vector $a^{-1}\xi a$, y éste, bajo la acción de la segunda, se convierte en el vector $\beta^{-1}(a^{-1}\xi a)\beta$. Por la ley asociativa, el último resultado se puede representar en la forma

$$\beta^{-1}(a^{-1}\xi a)\beta = (\alpha\beta)^{-1}\xi\alpha\beta.$$

Puesto que la multiplicación de un vector —a saber, el cuatérno vectorial $\vec{\xi}$ — por el cuatérno $(\alpha\beta)^{-1}$ a la izquierda y por el cuatérno $\alpha\beta$ a la derecha es equivalente a la rotación de este vector en un cierto ángulo alrededor del correspondiente eje, llegamos a la conclusión de que el resultado de dos rotaciones sucesivas caracterizadas por los cuatérnos α y β es la rotación caracterizada por el producto $\alpha\beta$. En otros términos, a la suma de rotaciones corresponde el producto de los cuatérnos que las representan.

Aparte de aplicaciones geométricas y físicas, los cuatérnos han encontrado notables aplicaciones en la teoría de números. Entre una serie de trabajos sobre estas cuestiones hay que mencionar, en particular, los de Ju. V. Linnik.

§12. Álgebras asociativas

Definición general de álgebra (sistemas hipercomplejos)

Hemos definido los números hipercomplejos como cantidades para cuya descripción son necesarios varios números reales; en realidad, para ser más concretos hemos considerado los números hipercomplejos simplemente como sistemas de números reales. Sin embargo, este punto de vista es demasiado estrecho y en las investigaciones teóricas se ha ido imponiendo gradualmente la siguiente definición.

Un cierto sistema de cantidades S se dice que es un álgebra (o sistema hipercomplejo) sobre el cuerpo P si:

a. Para todo elemento a de P y toda cantidad α del sistema S está definido un elemento del sistema que se llama producto de a por α y se denota por $a\alpha$.

b. Para cada dos cantidades α, β del sistema, está definida unívocamente cierta cantidad del sistema llamada suma de las dos cantidades y que se denota por $\alpha + \beta$.

c. Dadas dos cantidades del sistema, existe otra unívocamente definida que se llama producto de ellas y se denota por $\alpha\beta$; y si estas tres operaciones tienen las siguientes propiedades ⁴:

$$1'. \quad \alpha + \beta = \beta + \alpha,$$

$$2'. \quad (\alpha + \beta) + \gamma = \alpha + (\beta + \gamma),$$

⁴ Denotamos por letras griegas las cantidades del sistema S , y por letras latinas los elementos del cuerpo P .

3'. El sistema S tiene una cantidad θ con la propiedad

$$\alpha + \theta = \alpha.$$

$$4'. \quad a(\alpha + \beta) = a\alpha + a\beta,$$

$$5'. \quad (\alpha + \beta)a = \alpha a + \beta a,$$

$$6'. \quad (ab)\alpha = a(b\alpha),$$

$$7'. \quad \theta\alpha = \theta, 1 \cdot \alpha = \alpha, \text{ donde } 1 \text{ es el elemento unidad del cuerpo } P.$$

8'. Entre las cantidades de S existen $\alpha_1, \alpha_2, \dots, \alpha_n$, tales que toda cantidad del sistema se puede expresar en función de ellas unívocamente en la forma $a_1\alpha_1 + a_2\alpha_2 + \dots + a_n\alpha_n$,

$$9'. \quad (a\alpha)\beta = a(a\beta) = a(\alpha\beta),$$

$$10'. \quad \alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma, (\beta + \gamma)\alpha = \beta\alpha + \gamma\alpha.$$

En esta definición los elementos del cuerpo arbitrario P juegan el mismo papel que hasta ahora han jugado los números reales. De la condición 8' se deduce claramente que toda cantidad hipercompleja está determinada por un sistema de n elementos $\alpha_1, \alpha_2, \dots, \alpha_n$ de P ; y que, por tanto, según la elección de P , puede venir determinada por n números complejos, n números reales, n números racionales, etcétera.

Los ocho primeros postulados significan que S forma un espacio lineal de dimensión finita (ver capítulo 15, §2) sobre el cuerpo P , al que llamaremos cuerpo de base del álgebra.

Las condiciones 9' y 10' se pueden combinar en forma de las ecuaciones siguientes:

$$(a\beta + b\gamma)\alpha = a(\beta\alpha) + b(\gamma\alpha),$$

$$\alpha(a\beta + b\gamma) = a(\alpha\beta) + b(\alpha\gamma),$$

de las que se sigue que la operación de multiplicar es lineal respecto a cada factor.

En los últimos años se ha dado preferencia al nombre «álgebra» frente al de «sistema hipercomplejo», ya que los elementos de «sistemas hipercomplejos» muy generales pueden diferir considerablemente de los números ordinarios, de modo que este nombre resulta inapropiado. Los términos «sistemas hipercomplejos» y «números hipercomplejos» se aplican ahora a las álgebras más simples, como, por ejemplo, los cuaternos ordinarios.

De las condiciones 1'-10' se deduce que en las álgebras no se requiere la asociatividad ni la conmutatividad de la multiplicación, ni la existencia de elemento unidad, ni la posibilidad de una «división».

Toda álgebra S tiene una base, es decir, un sistema de elementos $a_1, a_2, \dots, a_n$ en función de los cuales todo elemento de S se puede escribir de manera única en la forma $a_1 a_1 + a_2 a_2 + \dots + a_n a_n$ con coeficientes del cuerpo P (cuerpo de base). Toda álgebra tiene una infinidad de bases, pero el número de elementos de cada una de ellas es siempre el mismo y se llama rango del álgebra.

El sistema de números complejos, considerado como un álgebra sobre el cuerpo de los números reales, tiene una base formada por los números 1, i . Pero los pares de números 2, $3i$ y 1, $a + bi$ (a, b números reales, $b \neq 0$) pueden servir también de base.

Sea $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ la base de un álgebra sobre un cuerpo P . Por definición, todo elemento del álgebra se puede escribir de una sola manera en la forma

$$a = a_1 \epsilon_1 + a_2 \epsilon_2 + \dots + a_n \epsilon_n.$$

Si $\beta = b_1 \epsilon_1 + \dots + b_n \epsilon_n$ es otro elemento del álgebra, por las propiedades de 1' a 6' tenemos

$$a + \beta = (a_1 + b_1) \epsilon_1 + (a_2 + b_2) \epsilon_2 + \dots + (a_n + b_n) \epsilon_n.$$

De modo similar, para todo a de P tenemos

$$aa = aa_1 \epsilon_1 + aa_2 \epsilon_2 + \dots + aa_n \epsilon_n.$$

Por tanto, la operación de suma de cantidades de un álgebra y la multiplicación por un elemento del cuerpo P están unívocamente determinadas por las fórmulas dadas. La operación de multiplicar cantidades de un álgebra se debe definir de modo expreso para cada álgebra; pero no necesitamos conocer la forma de multiplicar cantidades arbitrarias, ya que con las leyes de multiplicar los elementos de la base es suficiente. Efectivamente, por las propiedades 9' y 10'

$$(a_1 \epsilon_1 + a_2 \epsilon_2 + \dots + a_n \epsilon_n)(b_1 \epsilon_1 + b_2 \epsilon_2 + \dots + b_n \epsilon_n) = \sum a_i b_j \cdot \epsilon_i \epsilon_j.$$

Cada uno de los productos $\epsilon_i \epsilon_j$ es una cierta cantidad del álgebra y se puede expresar, por tanto, en función de los elementos de la base

$$\epsilon_i \epsilon_j = c_{ij1} \epsilon_1 + c_{ij2} \epsilon_2 + \dots + c_{ijn} \epsilon_n.$$

Aquí c_{ijk} denota un elemento del cuerpo de base P sobre el que se construye el álgebra. El primer índice denota el número del primer factor; el segundo, el número del segundo, y el tercero indica el número del elemento del cual es coeficiente. Los coeficientes c_{ijk} se llaman constantes de estructura del álgebra, ya que si conocemos estas constantes podemos efectuar cualquier operación con las cantidades del álgebra.

Es fácil saber el número de constantes de estructuras de un álgebra de rango n . Toda constante tiene tres índices, i, j, k . Por tanto, para un álgebra de rango n este número es el número de ternas que se pueden formar con los números $1, 2, \dots, n$, es decir, igual a n^3 . Por ejemplo, el sistema de los números complejos sobre el cuerpo de los números reales tiene una base formada por los dos números $1, i$. En virtud de las ecuaciones

$$\begin{aligned} 1 \cdot 1 &= 1 \cdot 1 + 0 \cdot i, & i \cdot 1 &= 0 \cdot 1 + 1 \cdot i, \\ 1 \cdot i &= 0 \cdot 1 + 1 \cdot i, & i \cdot i &= -1 \cdot 1 + 0 \cdot i, \end{aligned}$$

las constantes de estructura son, respectivamente, iguales a

$$\begin{aligned} c_{111} &= 1, & c_{112} &= 0, & c_{211} &= 0, & c_{212} &= 1, \\ c_{121} &= 0, & c_{122} &= 1, & c_{221} &= -1, & c_{222} &= 0. \end{aligned}$$

Supongamos, inversamente, n^3 elementos c_{ijk} de un cuerpo arbitrario P , cuyos índices son ternas de números naturales ($i, j, k = 1, 2, \dots, n$). Podemos tomar dichos elementos como constantes de estructura de un álgebra sobre el cuerpo P , utilizando la ecuación $\epsilon_i \epsilon_j = \sum_{k=1}^n c_{ijk} \epsilon_k$ como definición de la operación de multiplicar sobre el álgebra.

Anteriormente hemos visto que toda álgebra tiene, en general, un número infinito de bases distintas. Las constantes de estructura dependen de la elección de la base, y por tanto sistemas distintos de constantes de estructura pueden definir la misma álgebra.

¿Qué álgebras se deben considerar distintas y cuáles no? En la teoría de álgebras es conveniente considerar iguales dos álgebras sobre el mismo cuerpo P si son isomorfas, esto es, si los elementos de una de ellas se pueden poner en correspondencia uno a uno con los de la otra, de forma que a la suma y producto de cada dos cantidades del primer álgebra le correspondan la suma y el producto de las cantidades correspondientes en la segunda, y que al producto

de un elemento de P por un elemento de la primera corresponda el producto del mismo elemento de P por la cantidad correspondiente en la segunda.

Esta definición muestra que en la teoría de álgebras se estudian solamente aquellas propiedades de las cantidades y sistemas de cantidades del álgebra que encuentran expresión en forma de propiedades de las tres operaciones básicas. Para decirlo brevemente, estudia propiedades de las operaciones realizadas con cantidades del álgebra, sin tener en cuenta la naturaleza de dichas cantidades.

Es fácil demostrar que, si dos álgebras son isomorfas, las cantidades que forman una base en una de las álgebras corresponden a cantidades que forman base en la otra, y que las constantes de estructura calculadas respecto a bases correspondientes son iguales. Inversamente, si dos álgebras tienen las mismas constantes de estructura (álgebras sobre el mismo cuerpo) en base conveniente, entonces tales álgebras son isomorfas.

Entre todas las álgebras, las asociativas —esto es, las álgebras en las que la operación de multiplicar satisface la propiedad asociativa— han desempeñado y aún desempeñan un papel muy importante. En la presente sección haremos una exposición de las propiedades de tales álgebras. Entre las álgebras no asociativas las más importantes son las álgebras de Lie, en las cuales la operación de multiplicar satisface las siguientes propiedades:

$$a\beta = -\beta a, \quad a(\beta\gamma) + \beta(\gamma a) + \gamma(a\beta) = 0.$$

Tienen interés debido a la estrecha relación que existe entre ellas y los grupos de Lie, que fueron discutidos en el §7.

El álgebra de matrices

Hemos señalado anteriormente que en el primer período del desarrollo de la teoría de sistemas hipercomplejos la atención se centró principalmente en el estudio de varios sistemas que por una u otra razón eran de interés para los investigadores. Hemos examinado ya algunos de estos sistemas. El estudio del álgebra de matrices, que desempeña un papel fundamental en la teoría general de álgebras, comenzó aproximadamente a mediados del siglo pasado. Recordemos brevemente las definiciones de las operaciones con matrices (ver capítulo 15, §1).

Una matriz sobre un cuerpo P es una colección de elementos del cuerpo colocados en forma de tabla rectangular. Dos matrices se

dicen iguales si los elementos de los mismos lugares son iguales. Aquí consideraremos solamente matrices cuadradas, en las que el número de filas es igual al de columnas. El número de filas o columnas de una matriz cuadrada se llama orden de la matriz.

Para sumar matrices del mismo orden se suman los elementos de los lugares correspondientes. El producto de una matriz por un número del cuerpo P es, por definición, una matriz cuyos elementos son los productos de los elementos de la matriz dada por este número. La operación de multiplicar matrices entre sí se define de una forma más complicada: el producto de dos matrices de orden n es otra matriz de orden n en la que el elemento de la fila i y la columna j es igual a la suma de los productos de los elementos de la fila i de la primera matriz por los correspondientes de la columna j de la segunda. Por ejemplo:

$$\begin{bmatrix} a & b \\ a_1 & b_1 \end{bmatrix} \begin{bmatrix} x & y \\ x_1 & y_1 \end{bmatrix} = \begin{bmatrix} ax + bx_1 & ay + by_1 \\ a_1x + b_1x & a_1y + b_1y_1 \end{bmatrix}.$$

El motivo para la elección de esta definición de multiplicación fue expuesto en el capítulo 15.

En virtud de las definiciones dadas, las matrices de orden n con elementos de un cuerpo P forman un sistema de cantidades que se pueden sumar, multiplicar por un elemento de P y multiplicar entre sí. Un cálculo directo muestra que se cumplen las propiedades 1'-10' que definen un álgebra. Además, es fácil demostrar que el producto de matrices verifica la propiedad asociativa. Por tanto, el sistema de las matrices cuadradas de un cierto orden n con elementos de un cuerpo P es un álgebra asociativa sobre este cuerpo.

La ecuación

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = a \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + b \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + c \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + d \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

muestra que las matrices del segundo miembro forman una base del álgebra de matrices de orden 2. De forma más general, si denotamos por ϵ_{ij} la matriz que tiene un 1 en la fila i y un 1 en la columna j , siendo ceros los números de los restantes lugares, tenemos la siguiente ecuación

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ \cdot & & \cdot \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \sum_{i,j} a_{ij} \epsilon_{ij},$$

que muestra que las matrices ϵ_{ij} forman una base del álgebra de matrices de orden n . Puesto que el número de matrices ϵ_{ij} es igual a n^2 , el rango del álgebra de matrices es también igual a n^2 . La tabla de multiplicar para los elementos de la base tiene la forma

$$\epsilon_{ij} \cdot \epsilon_{kl} = \epsilon_{il}, \quad \epsilon_{ij} \cdot \epsilon_{kl} = 0, \quad j \neq k, \quad i, j, k, l = 1, 2, \dots, n.$$

El álgebra de matrices contiene un elemento unidad que es la matriz unidad.

Representaciones de álgebras asociativas

Supongamos que se asocia a cada cantidad de un álgebra A sobre un cuerpo P otra cierta cantidad de un álgebra B sobre el mismo cuerpo P . Si a la suma y el producto de cada dos elementos de A se asocian la suma y el producto de los correspondientes de B , y al producto de cada elemento de P por un elemento arbitrario de A se le asocia el producto del mismo elemento de P por el correspondiente de A , decimos que el álgebra A es homomorfa al álgebra B . Un homomorfismo de un álgebra asociativa A en el álgebra de matrices de orden n se llama una representación de A de grado n . Si a elementos distintos de A corresponden matrices distintas, la representación se llama fiel o isomorfa. Cuando un álgebra A viene representada isomórficamente por matrices, podemos suponer que las operaciones sobre las cantidades del álgebra se reducen a operaciones sobre las matrices correspondientes. Por tanto, la tarea de encontrar representaciones de álgebras reviste mucho interés. Aquí consideraremos solamente los métodos más sencillos de hallar tales representaciones; sin embargo, estos métodos juegan un importante papel en la teoría general.

Tomemos una base arbitraria $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ en un álgebra asociativa A , y sea α una cantidad arbitraria de A . Los productos $\epsilon_1\alpha, \epsilon_2\alpha, \dots, \epsilon_n\alpha$ son cantidades de A , y pueden por tanto expresarse en función de los elementos $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ de la base. Supongamos que

$$\begin{aligned} \epsilon_1\alpha &= a_{11}\epsilon_1 + a_{12}\epsilon_2 + \dots + a_{1n}\epsilon_n, \\ \epsilon_2\alpha &= a_{21}\epsilon_1 + a_{22}\epsilon_2 + \dots + a_{2n}\epsilon_n, \\ &\dots\dots\dots \\ \epsilon_n\alpha &= a_{n1}\epsilon_1 + a_{n2}\epsilon_2 + \dots + a_{nn}\epsilon_n. \end{aligned}$$

Como vemos, para una base fija podemos asociar a cada elemento a de A una matriz concreta $||a_{ij}||$. Un cálculo sencillo muestra que esta correspondencia es una representación de A , que suele llamarse representación regular de A . Su grado es evidentemente igual al rango del álgebra.

Los números complejos cabe considerarlos como un álgebra de rango 2 sobre el cuerpo de los números reales con la base $1, i$. Las ecuaciones

$$\begin{aligned}1 \cdot (a + bi) &= a \cdot 1 + b \cdot i, \\ i \cdot (a + bi) &= -b \cdot 1 + a \cdot i\end{aligned}$$

muestran que en la correspondiente representación regular el número complejo $a + bi$ corresponde a la matriz

$$\begin{bmatrix} a & b \\ -b & a \end{bmatrix}.$$

La representación análoga para cuaternos tiene la forma

$$a + bi + cj + dk \rightarrow \begin{bmatrix} a & b & c & d \\ -b & a & -d & c \\ -c & d & a & -b \\ -d & -c & b & a \end{bmatrix}.$$

Estas representaciones de los complejos y los cuaternos son fieles (es decir, son representaciones isomorfas a las álgebras). Hay ejemplos que muestran, sin embargo, que la representación regular no siempre es fiel. Pero si el álgebra contiene un elemento unidad, su representación regular es necesariamente fiel.

Es fácil probar que toda álgebra asociativa se puede sumergir en otra con elemento unidad. La representación regular de esta última es fiel; por consiguiente, esta representación del álgebra dada es también fiel. Así, pues, toda álgebra asociativa tiene una representación fiel por medio de matrices.

Este método de construir representaciones es insuficiente para hallar todas las representaciones de un álgebra. Existe otro método más fino que está relacionado con el concepto de «ideal». El concepto de ideal juega un importante papel en las matemáticas modernas.

Un sistema I de elementos de un álgebra recibe el nombre de *ideal* a la derecha si es un subespacio lineal del álgebra y si el producto a la derecha de todo elemento de I por cualquier elemento del

Supongamos que un álgebra asociativa A contiene un ideal a la derecha I . Tomemos en él una base $\epsilon_1, \epsilon_2, \dots, \epsilon_m$. Puesto que en general I es sólo una parte de A , sus bases tendrán en general menos elementos que las de A . Sea α un elemento de A . Como I es un ideal a la derecha y $\epsilon_1, \epsilon_2, \dots, \epsilon_m$ están contenidos en I , los productos $\epsilon_1\alpha, \dots, \epsilon_m\alpha$ pertenecen también a I y se pueden expresar, por tanto, linealmente en función de la base $\epsilon_1, \dots, \epsilon_m$; es decir,

$$\begin{aligned}\epsilon_1\alpha &= a_{11}\epsilon_1 + \cdots + a_{1m}\epsilon_m \\&\dots\dots\dots\\ \epsilon_m\alpha &= a_{m1}\epsilon_1 + \cdots + a_{mm}\epsilon_m.\end{aligned}$$

La estructura de las álgebras

Aunque este resultado es en principio de interés, ya que reduce el problema de hallar todas las álgebras al de encontrar todas las posibles subálgebras de las álgebras de matrices, no da una respuesta

directa al problema de la estructura de las álgebras. La primera respuesta general a este problema fue dada a finales del pasado siglo por F. E. Molin (1861-1941), profesor de la Universidad de Dorpat (Tartu), que enseñó en el Instituto Politécnico de Tomsk alrededor de 1900.

Un álgebra se llama *simple* si no contiene más ideales biláteros que el ideal cero y el álgebra completa. Molin probó que toda álgebra asociativa simple de rango 2 o más de 2 sobre el cuerpo de los números complejos es isomorfa al álgebra de todas las matrices de un orden conveniente sobre este cuerpo.

Continuando las investigaciones de Molin, Wedderburn obtuvo a principios del siglo xx diversos resultados que dan una descripción muy completa de la estructura de las álgebras sobre un cuerpo arbitrario.

Un sistema arbitrario de elementos de un álgebra A (en particular la propia álgebra es o un ideal o una subálgebra de sí misma) se llama *nihilpotente* si existe un número natural s tal que la potencia de exponente s de cualquier elemento del sistema es igual a cero. Toda álgebra asociativa tiene un único ideal maximal bilátero nihilpotente, que se llama radical del álgebra. Un álgebra cuyo radical es igual a cero se llama *semisimple*. Se puede demostrar que toda álgebra semisimple se divide en una suma de una clase especial de álgebras simples; en virtud de ello, el estudio de las álgebras semisimples se reduce por completo al de las simples. Por último, un álgebra A se llama *álgebra de división* si toda ecuación de la forma $ax = b$ ($a \neq 0$) tiene solución en A .

La estructura de las álgebras simples sobre el cuerpo de los números complejos está completamente descrita por el teorema de Molin citado anteriormente. Pero si el cuerpo de base P es arbitrario, entonces se cumple el teorema de Wedderburn, que es más general: toda álgebra simple de rango 2 o superior sobre un cuerpo P es isomorfa al álgebra de todas las matrices de un orden conveniente con elementos de una cierta álgebra de división sobre el mismo cuerpo P . De este modo, el teorema de Wedderburn reduce el problema de encontrar álgebras simples sobre un cuerpo P al de hallar álgebras de división sobre P . Sobre el cuerpo de los complejos existe una sola álgebra con división, que es el propio cuerpo de los complejos. Se sigue de aquí, por el teorema de Wedderburn, que todas las álgebras simples sobre el cuerpo de los complejos son isomorfas a un álgebra de matrices sobre el mismo cuerpo, es decir, el teorema de Molin.

Sobre el cuerpo de los números reales existen solamente tres álgebras asociativas con división: el cuerpo de los números reales, el cuerpo de los números complejos y el álgebra de cuaternos. La demostración de este teorema no es muy fácil, y la omitimos. Por el teorema de Wedderburn esto implica que toda álgebra simple sobre el cuerpo de los números reales es isomorfa al álgebra de matrices de un orden conveniente, bien sobre el cuerpo de los reales, bien sobre el cuerpo de los números complejos o bien sobre el álgebra de cuaternos.

De estos ejemplos se deduce cómo se pueden describir las álgebras semisimples a partir de los teoremas de Molin y Wedderburn. En cuanto a las álgebras con un radical, es de gran importancia el teorema fundamental de Wedderburn; de acuerdo con este teorema, imponiendo ciertas restricciones al cuerpo de base, toda álgebra A con un radical R tiene un subálgebra semisimple L tal que todo elemento de A se puede representar unívocamente en forma de suma $\lambda + \varrho$ ($\lambda \in L$, $\varrho \in R$), donde la subálgebra L está en un cierto sentido unívocamente determinada dentro de A .

Los teoremas fundamentales que acabamos de formular dan una idea de los posibles tipos de álgebras asociativas y reducen la cuestión de su estructura esencialmente al problema de la estructura de álgebras nilpotentes. La teoría de estas últimas está aún en período de desarrollo.

§13. *Algebras de Lie*

En la sección 12 dijimos que, además de la teoría de las álgebras asociativas, la teoría de álgebras de Lie ha sido elaborada con gran detalle en estos últimos años; en ellas, la multiplicación está sometida a las reglas siguientes:

$$\alpha\beta = -\beta\alpha, \quad \alpha(\beta\gamma) + \beta(\gamma\alpha) + \gamma(\alpha\beta) = 0.$$

La importancia de estas álgebras se debe a que están íntimamente relacionadas con los grupos de Lie (ver §7), es decir, la clase más importante de grupos continuos. Como ya hemos dicho, los grupos de Lie ocupan un importante lugar en la geometría contemporánea. Debido al origen de los grupos y álgebras de Lie, el mayor interés recae en las álgebras de Lie sobre el cuerpo de los números reales o los complejos.

Uno de los ejemplos más sencillos de álgebras de Lie es el siguiente. Consideremos el conjunto de todas las matrices cuadradas de orden n arbitrario. Introducimos en ellas una operación de conmutación; entendemos por esto la formación del llamado conmutador $AB - BA$ de dos matrices A y B , que se denota por $[A, B]$.

Es fácil verificar que

$$[A, B] = -[B, A],$$

$$[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0.$$

En consecuencia, el conjunto de las matrices cuadradas de un cierto orden dado n forman un álgebra de Lie para la operación de conmutación. Está claro que toda subálgebra del álgebra de Lie formada por matrices, esto es, todo conjunto de matrices que es cerrado respecto a la suma, producto por un número del cuerpo de base y conmutación, es a su vez un álgebra de Lie.

La cuestión de si para toda álgebra de Lie definida abstractamente existe un álgebra de matrices isomorfa a ella permaneció abierta durante mucho tiempo. Fue resuelta afirmativamente sólo en 1935 por I. D. Ado, discípulo del famoso algebrista H. G. Chebotarev.

Esbozemos ahora en términos generales, sin entrar en detalles ni dar resultados rigurosos, las relaciones que existen entre los grupos de Lie y las álgebras de Lie, restringiéndonos al caso en que ambos están representados por matrices.

Sea L una cierta álgebra de Lie de matrices. A toda matriz A perteneciente a L asociamos la matriz $U = e^A = E + A/1! + A^2/2! + \dots$. La colección de todas las matrices obtenidas por este método forman un grupo de Lie para la multiplicación ordinaria de matrices. Inversamente, para todo grupo de Lie podemos encontrar una única álgebra de Lie (salvo isomorfismos) tal que el grupo correspondiente a ésta es precisamente el grupo dado (salvo isomorfismos). Para mayor sencillez hemos dado una formulación simplificada del teorema sobre la conexión entre álgebras y grupos de Lie. En realidad, la relación $U = e^A$ existe sólo para U y A suficientemente «próximas» a las matrices unidad y cero respectivamente. Una formulación rigurosa requeriría la introducción de los conceptos más complicados de grupo local e isomorfismo local.

De este modo, el paso del álgebra de Lie al grupo correspondiente se efectúa mediante una operación similar a la exponenciación; y el paso inverso, del grupo al álgebra, por una operación similar a la de tomar logaritmos.

Si L coincide con el álgebra de todas las matrices de orden n , entonces el grupo de Lie correspondiente es el grupo de todas las matrices no singulares, ya que toda matriz U próxima a la matriz unidad se puede representar en la forma $U = e^A$.

Una matriz $A = ||A_{ij}||$ se llama antisimétrica si sus elementos satisfacen la relación $a_{ji} = -a_{ij}$. Las matrices antisimétricas forman un álgebra de Lie, porque si A y B son antisimétricas también lo son $AB - BA = [A, B]$ y $\alpha A + \beta B$.

Es fácil verificar que para toda matriz antisimétrica A la expresión e^A es una matriz ortogonal y que toda matriz ortogonal próxima a la matriz unidad se puede representar en forma exponencial. Por consiguiente el álgebra de Lie del grupo de matrices ortogonales es el álgebra de matrices antisimétricas.

Por geometría analítica sabemos que todo giro del espacio de centro el origen de coordenadas viene dado por una matriz ortogonal y que al producto de giros corresponde el producto de las matrices correspondientes. En otros términos, el grupo de rotaciones del espacio alrededor de un cierto punto fijo es isomorfo al grupo de matrices ortogonales de orden 3. Se deduce de aquí que el álgebra de Lie para el grupo de rotaciones del espacio es el álgebra de todas las matrices antisimétricas de orden 3, esto es, el álgebra de Lie de matrices de la forma

$$A = \begin{bmatrix} 0 & -a & -b \\ a & 0 & -c \\ b & c & 0 \end{bmatrix}.$$

Puesto que cada una de estas matrices está caracterizada por tres números, a, b, c , se las puede representar por vectores en el espacio cuyas proyecciones sobre los ejes coordenados sean a, b, c . Una combinación lineal de las matrices A_1 y A_2 , $\alpha A_1 + \beta A_2$, está asociada evidentemente con la combinación lineal de los vectores correspondientes $\alpha a_1 + \beta a_2$, y el conmutador de las matrices

$$\begin{aligned} [A_1, A_2] &= A_1 A_2 - A_2 A_1 = \\ &= \begin{bmatrix} 0 & -a_1 & -b_1 \\ a_1 & 0 & -c_1 \\ b_1 & c_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -a_2 & -b_2 \\ a_2 & 0 & -c_2 \\ b_2 & c_2 & 0 \end{bmatrix} - \begin{bmatrix} 0 & -a_2 & -b_2 \\ a_2 & 0 & -c_2 \\ b_2 & c_2 & 0 \end{bmatrix} \begin{bmatrix} 0 & -a_1 & -b_1 \\ a_1 & 0 & -c_1 \\ b_1 & c_1 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} 0 & b_2 c_1 - b_1 c_2 & a_1 c_2 - a_2 c_1 \\ b_1 c_2 - b_2 c_1 & 0 & a_2 b_1 - a_1 b_2 \\ a_2 c_1 - a_1 c_2 & a_1 b_2 - a_2 b_1 & 0 \end{bmatrix} \end{aligned}$$

está asociado con el vector cuyas componentes son $b_1c_2 - b_2c_1$, $a_2c_1 - a_1c_2$, $a_1b_2 - a_2b_1$; es decir, con el vector producto de $\mathbf{a}_1$ y $\mathbf{a}_2$. Hemos llegado al notable resultado que afirma que el conjunto de los vectores ordinarios para las operaciones de suma, multiplicación por un escalar y producto de vectores forma un álgebra de Lie que corresponde al grupo de los giros del espacio de centro un punto fijo. Esto muestra, al mismo tiempo, cuán íntimamente ligados están los conceptos geométricos con el grupo de giros en el espacio, o dicho en otros términos: con las leyes de movimiento de los cuerpos rígidos.

A finales del siglo pasado y comienzos del presente se obtuvieron una serie de resultados para álgebras de Lie que son similares a los resultados fundamentales de las álgebras asociativas, aunque las demostraciones y enunciados son más complicados. Así, como resultado de los esfuerzos de Lie, Killing y Cartan se establecieron con éxito los conceptos de radical y semisimplicidad para un álgebra de Lie y se encontraron todas las álgebras de Lie simples sobre los cuerpos de los números reales y complejos. A principios de los años treinta, Cartan y Weyl construyeron la teoría de representación de álgebras de Lie mediante matrices, que se convirtió en un poderoso instrumento para la resolución de muchos problemas. En los últimos quince años varios matemáticos soviéticos han participado en el desarrollo de la teoría de álgebras de Lie, alcanzando resultados notables. En particular, realizaron importantes progresos en la teoría de representaciones y dieron soluciones definitivas a los problemas de subálgebras semisimples de álgebras de Lie, de la estructura de álgebras con un radical dado, y otros.

§14. Anillos

En la sección 11 hemos dado la definición general de cuerpo como un conjunto arbitrario de elementos en el que las operaciones de suma y multiplicación satisfacen las propiedades 1 a 10. Omitiendo en esta definición el postulado 10 sobre la existencia de una división, así como los postulados 7 y 8 sobre la conmutatividad y asociatividad de la multiplicación, obtenemos la definición del concepto de anillo, uno de los más importantes del álgebra contemporánea.

Todo cuerpo, y también toda álgebra, considerados solamente respecto a las operaciones de suma y multiplicación es un anillo. Un ejemplo aún más sencillo de anillo lo constituyen los números enteros con las operaciones usuales de suma y multiplicación de

números. Para las mismas operaciones, los conjuntos de números de la forma $a + bi$, $a + b\sqrt{2}$, $a + b\sqrt[3]{2} + c\sqrt[3]{4}$, etc., son también anillos, donde a, b, c son enteros. Los elementos de estos anillos son números y los anillos se llaman por tanto anillos de números. En el capítulo 4 y en el 10 se discutieron algunas propiedades importantes de estos anillos.

Sin embargo existen importantes clases de anillos cuyos elementos no son números. Por ejemplo, el conjunto de polinomios de varias variables $x_1, x_2, \dots, x_n$ con coeficientes en un anillo o cuerpo es, para las operaciones de suma y producto de polinomios, otro anillo; igualmente lo son el conjunto de las funciones continuas definidas en un cierto dominio, el conjunto de las transformaciones lineales de un espacio lineal o de Hilbert, etc.

Las propiedades aritméticas de los anillos numéricos constituyen el objeto de la profunda teoría de números algebraicos, que se encuentra a caballo entre el álgebra propiamente dicha y la teoría de números. El estudio de las propiedades de los anillos de polinomios es el objeto de la llamada teoría de ideales de polinomios, que está conectada con las ramas superiores de la geometría analítica. Por último, los anillos de funciones y transformaciones juegan un importante papel en análisis funcional (ver capítulo 18).

Sobre la base de esta y otras teorías concretas se ha desarrollado en el presente siglo la teoría general de anillos y la de anillos topológicos.

Por razones de espacio daremos solamente los rudimentos de la teoría de anillos.

Ideales

Un subconjunto I de elementos de un anillo K (no necesariamente asociativo) se dice que es un *ideal* si la diferencia de dos elementos cualesquiera de I pertenece a I y si los productos ax, xa , donde a pertenece a I , y x a K , pertenecen también a I cualesquiera que sean a y x .

Todo ideal de un anillo es también un anillo para las operaciones de suma y multiplicación definidas en el anillo. Tales partes se llaman subanillos del anillo dado y por consiguiente todo ideal es al mismo tiempo un subanillo. La proposición inversa no es cierta en general.

La intersección de un sistema arbitrario de ideales es otro ideal; en particular, la intersección de todos los ideales que contienen un elemento fijo a del anillo es un ideal que recibe el nombre de ideal principal engendrado por el elemento a y se denota por (a) .

El concepto de ideal engendrado por dos o más elementos se define de forma similar. Es fácil ver que, si un anillo asociativo y conmutativo tiene elemento unidad, el ideal engendrado por los elementos $a_1, a_2, \dots, a_n$ es simplemente la colección de todos los elementos del anillo que admiten una representación de la forma $x_1a_1 + \dots + x_na_n$, donde $x_1, \dots, x_n$ son elementos arbitrarios del anillo. En particular, el ideal principal (a) en un anillo conmutativo y asociativo con elemento unidad es la colección de los elementos de la forma xa , es decir, de múltiplos de a .

En el anillo de los enteros todo anillo es principal. El anillo de los polinomios en una variable con coeficientes en un cuerpo cualquiera tiene la misma propiedad, e igualmente el anillo de los complejos de la forma $a + bi$, donde a y b son enteros, así como otros. Sin embargo, el conjunto de los polinomios en dos variables x, y sin término independiente es un ideal, pero no principal, en el anillo de todos los polinomios en x e y con coeficientes racionales.

De la misma forma que hicimos para un subgrupo normal en la teoría de grupos, podemos construir para todo ideal I de un anillo K un anillo de clases residuales (o anillo factor) K/I . Esto se hace como sigue. Dos elementos a, b de K se dicen congruentes módulo el ideal I —simbólicamente $a \equiv b(I)$ — si su diferencia está contenida en I . Se comprueba fácilmente que la relación de congruencia es reflexiva, simétrica y transitiva (ver capítulo 14), de modo que todos los elementos de K se distribuyen en clases de elementos congruentes (módulo I). Si consideramos ahora estas clases como elementos de un nuevo conjunto, podemos introducir para ellos el concepto de suma y producto: la «suma» de dos clases será la clase que contiene la suma de dos elementos (representantes), uno de cada clase; y el producto, la clase que contiene el producto de dos representantes. De la definición de ideal se sigue que la suma y el producto no dependen de la elección de los representantes y que, como resultado de ello, el conjunto de las clases es un anillo.

El papel del anillo de clases residuales es, en la teoría de anillos, enteramente análogo al del grupo factor en la teoría de grupos. En particular, la construcción del anillo de clases residuales es un método conveniente para formar nuevos anillos con propiedades diversas. Además, es fácil de ver, por ejemplo, que un anillo arbitrario conmutativo es isomorfo al anillo factor de un anillo de polinomios con coeficientes enteros y en un número de variables suficientemente grande.

Propiedades aritméticas de los anillos

En los anillos de números y en los cuerpos el producto de varios elementos sólo puede ser cero si al menos uno de ellos lo es. En anillos arbitrarios esto no tiene por qué ser cierto: por ejemplo, el producto de dos matrices no nulas puede ser nulo. Si en cierto anillo $ab = 0$ y $a \neq 0$, $b \neq 0$, entonces a y b se llaman divisores de cero. Si un anillo no tiene tales elementos se dice que es un *anillo sin divisores de cero*.

Para el estudio de las leyes de divisibilidad en anillos suponemos habitualmente que son conmutativos y que no tienen divisores de cero. Tales anillos se llaman a veces dominios de integridad. Los anillos numéricos y de polinomios citados anteriormente son dominios de integridad.

Sea K un dominio de integridad. Decimos que un elemento a es divisible en K por el elemento b si $a = bq$, $q \in K$. De esta definición se sigue que la suma de elementos divisibles por b también lo es, y el producto de varios factores es necesariamente divisible por b si uno de ellos lo es. Cuando se intenta introducir en la teoría de anillos el concepto de elemento primo llegamos a complicaciones, como ya mencionamos en el capítulo 10. Para empezar tenemos que introducir el concepto de elementos asociados de un anillo, llamando asociados a los elementos a, b si a es divisible por b y b divisible por a . Escribiendo $a = bq_1$, $b = aq_2$, tenemos $ab = ab \cdot q_1q_2$; es decir, $q_1q_2 = e$, donde e es el elemento unidad de K . Los cocientes de elementos asociados se llaman por tanto divisores de uno o unidades. Todo elemento del dominio es divisible por cada unidad. En el anillo de los enteros las unidades son ± 1 , en el anillo de los números de la forma $a + bi$, con a, b enteros, las unidades son los números $\pm 1, \pm i$.

Todo elemento de un dominio de integridad K se puede descomponer en la forma $a = a\epsilon \cdot \epsilon^{-1}$, donde ϵ es una unidad cualquiera. Estas descomposiciones se llaman triviales. Si a no tiene otras descomposiciones, si dice que a es primo o que es un elemento indescomponible en K . En relación con el importante teorema sobre la unicidad de la descomposición de los enteros en factores primos, es interesante encontrar tales clases de anillos —y entre ellos los no conmutativos— en los que se cumpla un teorema similar. Este teorema se presenta, por ejemplo, en los anillos principales, esto es, en los dominios de integridad en los que todos los ideales son principales.

El propio concepto de ideal surgió en relación con el problema de la unicidad de la descomposición en factores primos. Aproxima-

damente a mediados del pasado siglo el matemático alemán Kummer, tratando de probar la famosa proposición de Fermat de que la ecuación $x^n + y^n = z^n$ no tiene soluciones enteras para $n \geq 3$, tuvo la idea de considerar números de la forma $a_0 + a_1\zeta + \cdots + a_n\zeta^{n-1}$, donde $\zeta = \cos 2\pi/n + i \operatorname{sen} 2\pi/n$ es una solución de la ecuación $x^n = 1$ y $a_0, \dots, a_n$ son enteros. Los números de este tipo forman un dominio de integridad, y Kummer, al principio, dio por supuesto que evidentemente se verificaba el teorema sobre la unicidad de la descomposición en factores primos en este dominio. Sobre esta base construyó una demostración del teorema de Fermat. Sin embargo, al repasar sus razonamientos observó que la suposición de la unicidad de la descomposición no es verdadera. Deseando conservar la unicidad de la descomposición en factores primos, Kummer se vio obligado a considerar descomposiciones de números del dominio en factores que no entran en el propio dominio. Llamó a estos números ideales. Posteriormente, en la construcción de la teoría general, los matemáticos introdujeron, en lugar de los números ideales, conjuntos de elementos del dominio que son divisibles por un número ideal u otro, y estos conjuntos fueron llamados ideales. El descubrimiento de la no unicidad de la descomposición en factores primos en anillos numéricos es uno de los hechos más interesantes descubiertos en el último siglo y condujo a la creación de la extensa teoría de números algebraicos.

Una de las aplicaciones más sorprendentes de este teorema al problema de la descomposición de enteros ordinarios en una suma de cuadrados fue mencionada al final del capítulo 10. El trabajo de los matemáticos E. I. Zolotarev, G. F. Voronoi, I. M. Vinogradov y N. G. Chebotarev desempeñó un importante papel en el desarrollo de la teoría de anillos numéricos.

Variedades algebraicas

La geometría algebraica es otra de las fuentes de la teoría de ideales. Cuando uno se acerca por primera vez a la teoría de curvas de segundo orden suele comprobar con asombro que con el nombre unitario de hipérbola se designa en realidad una colección de dos curvas no conexas, a saber las dos ramas de la hipérbola, y que un par de rectas recibe el nombre de curva degenerada de segundo orden. Esta cuestión de terminología se aclara en álgebra: si consideramos las ecuaciones de las curvas en la forma $f(x, y) = 0$, donde $f(x, y)$ es un polinomio en x, y , entonces en el primer caso el primer miembro de esta ecuación es un polinomio irreducible de segundo

grado, y en el segundo el producto de dos polinomios de primer grado. Una curva cuya ecuación se puede representar por medio de un polinomio irreducible $f(x, y)$ se llama irreducible, y en caso contrario, reducible.

Al pasar a curvas en el espacio, la cuestión se torna más complicada. Una curva en el espacio se puede representar por un sistema de dos ecuaciones $f(x, y, z) = 0$, $g(x, y, z) = 0$, donde los polinomios no están, en modo alguno, determinados unívocamente por la curva. ¿A qué llamaremos aquí una curva irreducible?

La respuesta natural la da en este caso la teoría de ideales. Sea $f_1, f_2, \dots$ un conjunto cualquiera de polinomios en las variables x, y, z con coeficientes complejos. El conjunto de todos los puntos del espacio (complejo) cuyas coordenadas hacen cero todos estos polinomios se dice que es la variedad algebraica definida por los polinomios dados. Denotemos por M esta variedad y consideremos todos los polinomios en las variables x, y, z que se anulan en todo punto de M . Es fácil ver que el conjunto I de todos estos polinomios es un ideal en el anillo de los polinomios en las variables x, y, z . Este ideal tiene además la propiedad de que, si una potencia de un polinomio pertenece a I , el polinomio también pertenece a él. Y sucede que, aunque la misma variedad algebraica se puede definir por distintos conjuntos de polinomios, la correspondencia entre variedades e ideales con la susodicha propiedad adicional es uno a uno.

Así, pues, al estudiar las propiedades de las variedades, es natural discutir no sus «ecuaciones», que son más o menos accidentales, sino los ideales correspondientes. Si un ideal I se puede representar en forma de intersección de otros dos ideales I_1, I_2 , entonces la variedad M es la unión de las variedades M_1, M_2 correspondientes a los ideales I_1, I_2 . De aquí que una variedad M se debe llamar, de modo natural, irreducible si su ideal correspondiente no es la intersección de otros dos que lo contienen. A la división de una curva en curvas de órdenes inferiores, a la descomposición de una variedad en otras irreducibles, corresponde ahora la representación de un ideal como intersección de otros no descomponibles. El problema de unicidad y posibilidad de tales descomposiciones es uno de los primeros en la teoría de variedades algebraicas y en la teoría general de ideales.

La estructura de los anillos no conmutativos

Toda álgebra es al mismo tiempo un anillo respecto a las operaciones de adición y multiplicación. Por tanto, muchos de los resultados y conceptos de la teoría de álgebras siguen siendo válidos para

anillos arbitrarios. Sin embargo, la demostración de resultados más profundos de las álgebras —como, por ejemplo, los teoremas de Molin y Wedderburn (ver §11)— en el caso de la teoría de anillos trae consigo grandes dificultades que sólo se han podido resolver parcialmente en los últimos diez o quince años. En primer lugar el problema de encontrar una definición tal de radical de un anillo que los anillos con radical cero tengan algún parecido con las álgebras semisimples y que todos los resultados de la teoría de estructura de álgebras se obtengan como casos especiales de los teoremas de la teoría de anillos. Existen hoy en la teoría de anillos una serie de definiciones de radical que permiten —con distintas restricciones en cada caso— construir una teoría satisfactoria para la estructura de anillos semisimples. Como hemos dicho anteriormente, el interés por la teoría de anillos no conmutativos está estimulado hasta cierto punto por el gran valor de la teoría de anillos de operadores en el análisis funcional.

§15. *Retículos*

Como el lector sabe, un conjunto de objetos se llama parcialmente ordenado si para ciertas parejas de sus elementos se puede determinar cuál de esos objetos precede o está subordinado al otro; se supone además que: 1) todo objeto se precede a sí mismo; 2) si a precede a b y b precede a a , se sigue que a y b son idénticos; 3) si a precede a b y b a c , se sigue que a precede a c . Esta relación de subordinación se denota usualmente por $\leq$.

Un ejemplo importante de conjunto parcialmente ordenado lo constituye la colección de todos los subconjuntos de un conjunto arbitrario, donde la relación de subordinación significa que un subconjunto es parte de otro.

Si la relación de subordinación está definida para todo par de elementos de un conjunto parcialmente ordenado, se dice que el conjunto está totalmente (o linealmente) ordenado. Conjuntos ordenados son, por ejemplo, los números reales, con la relación $a \leq b$ que significa que a no es mayor que b . Por el contrario, el conjunto parcialmente ordenado de todas las partes de una colección con más de un elemento no está totalmente ordenado, ya que los conjuntos sin elementos comunes no son comparables.

Supongamos que los elementos de un conjunto parcialmente ordenado M tienen la propiedad de que todo par a, b posee un extremo superior c , esto es, existe un elemento c tal que $a \leq c$, $b \leq c$ y tal que para todo elemento d de M que cumple las condiciones $a \leq d$,

$b \leq d$ tenemos $c \leq d$. Entonces M recibe el nombre de semirretículo superior y el elemento c es la «suma» de a y b . Es fácil verificar que esta «suma» verifica las siguientes propiedades:

$$a + b = b + a, \quad (a + b) + c = a + (b + c), \quad a + a = a. \quad (13)$$

Un hecho muy notable es que el inverso también se verifica. Si en un conjunto se define una operación con las propiedades (13), entonces, diciendo que a precede a b si $a + b = b$, obtenemos un conjunto parcialmente ordenado en el que $a + b$ es el extremo superior de a y b .

De modo similar podemos definir los semirretículos inferiores sin más que considerar extremos inferiores en lugar de superiores; el extremo inferior se suele llamar «producto» de los elementos dados. Esta operación tiene las mismas propiedades que la adición, a saber

$$ab = ba, \quad (ab)c = a(bc), \quad aa = a. \quad (14)$$

Un conjunto parcialmente ordenado que es al mismo tiempo retículo superior e inferior se llama *retículo*. Por lo que hemos dicho, en todo retículo se pueden definir dos operaciones sujetas a las condiciones (13) y (14). Sin embargo, estas operaciones están relacionadas una con la otra, ya que la relación $a \leq b$ en un retículo se puede escribir en cualquiera de las formas $a + b = b$, $ab = a$. Dicho con otras palabras, en un retículo las ecuaciones $a + b = b$ y $ab = a$ deben ser equivalentes. Las últimas ecuaciones se pueden escribir algebraicamente en forma de ecuaciones, así:

$$a + ab = a, \quad a(a + b) = a, \quad (15)$$

y, en virtud de ello, el estudio de retículos se convierte en una tarea puramente algebraica consistente en estudiar sistemas con las operaciones sometidas a las condiciones (13), (14) y (15). La importancia del enfoque algebraico al estudio de los retículos consiste, hablando sin rigor, en que las peculiaridades de uno u otro retículo concreto se pueden expresar convenientemente en forma de relaciones algebraicas de una u otra clase entre los elementos; también se puede sacar partido de la teoría clásica de grupos y anillos.

Como ya hemos dicho, la colección de todos los subconjuntos de un conjunto es un conjunto parcialmente ordenado. No es difícil comprobar que es un retículo con las operaciones de unión como suma

y de intersección como multiplicación. Si consideramos sólo una parte de los subconjuntos, obtenemos distintos retículos. Por ejemplo, son retículos el conjunto de todos los subgrupos y también el conjunto de todos los subgrupos normales de un grupo dado, el conjunto de todos los subanillos y el de todos los ideales de un anillo cualquiera, etcétera. En particular, en los retículos de todos los subgrupos normales de un grupo y de todos los ideales de un anillo se cumple, aparte de las relaciones (13), (14) y (15), también la llamada ley modular

$$a(ab + c) = ab + ac.$$

La teoría de retículos con la ley modular (retículos de Dedekind o modulares) es un capítulo importante en la teoría general de retículos.

Muchos de los teoremas de las teorías de grupos y anillos son afirmaciones acerca de la distribución de subgrupos, subgrupos normales, e ideales; por consiguiente, estos teoremas se pueden volver a formular como teoremas sobre retículos de subgrupos o ideales. Teoremas similares se verifican, con algunas restricciones, para retículos generales. Por este procedimiento se han trasladado a la teoría de retículos teoremas de la teoría de grupos, de la teoría de anillos y de otros campos. Por otra parte, la aplicación de los métodos de la teoría de retículos ha resultado muy útil para hallar propiedades de retículos concretos, por ejemplo, en la teoría de grupos y en la de anillos.

La teoría de retículos se ha desarrollado recientemente, durante los años veinte y treinta de nuestro siglo, y no ha encontrado aún aplicaciones tan importantes como, por ejemplo, la teoría de grupos. Sin embargo, hoy día constituye una disciplina bien formada, de rico contenido y con una amplia gama de problemas.

§16. *Otros sistemas algebraicos*

En las secciones precedentes hemos intentado explicar cómo la aplicación de los métodos algebraicos a una gama de problemas cada vez más amplia ha conducido a una extensión del sistema de objetos que se estudian en álgebra y a la generalización del concepto de operación algebraica. En este contexto, el método axiomático nacido del trabajo de I. N. Lobachevski sobre los fundamentos de la geometría y el desarrollo de la teoría general de conjuntos han jugado un importante papel.

Uno de los resultados fundamentales fue el gradual esclarecimiento de los conceptos generales de operación algebraica y sistema algebraico y la acumulación de los hechos más importantes referentes a la definición de sistemas algebraicos. El álgebra moderna parte del concepto general de operación, y no de operaciones definidas de un modo concreto para números, como se hace en el álgebra escolar. Supongamos un cierto sistema S de elementos y en él una regla que asocia a cada sistema de m elementos $a_1, a_2, \dots, a_m$, en un orden concreto, un elemento bien definido a del mismo sistema. Se dice entonces que se ha definido en S una operación m -aria y que a es el resultado de esta operación realizada sobre los elementos $a_1, a_2, \dots, a_m$. Un conjunto de elementos, junto con una o varias operaciones definidas sobre él, recibe el nombre de sistema algebraico. Una de las tareas más importantes del álgebra es el estudio y clasificación de estos sistemas. Sin embargo, formulado de esta forma el problema tiene un carácter demasiado general. De hecho, sólo ciertos sistemas han demostrado hasta el presente ser importantes y capaces de dar lugar a teorías interesantes. Por ejemplo, de los sistemas con una sola operación sólo la teoría de grupos, a la que hemos dedicado las secciones 1-10, ha dado lugar a una profunda ciencia matemática; y entre los de dos operaciones o más, los de mayor interés son los cuerpos, álgebras, anillos y retículos. Sin embargo, el número de sistemas que por una u otra razón se estudian hoy día crece continuamente. Al mismo tiempo, ciertas ramas clásicas de las matemáticas tales como, por ejemplo, el estudio de homomorfismos, de sistemas libres y uniones libres, de uniones directas, y recientemente el estudio de radicales, han sido extendidas con éxito a la teoría general de sistemas algebraicos.

Al discutir el carácter del álgebra como un todo, se subraya a menudo que la completa ausencia o el papel subordinado del concepto de continuidad es una de sus peculiaridades, de modo que el álgebra se considera como una ciencia con una cierta preferencia por lo discreto. Este juicio refleja indudablemente un rasgo objetivo importante del álgebra. En el mundo de lo real, la continuidad y la discontinuidad se encuentran en unidad dialéctica. Pero para conocer la realidad es a menudo necesario dividirla en partes y estudiarlas separadamente. Por consiguiente, la atención unilateral que el álgebra presta a las relaciones discretas no debe considerarse como una deficiencia.

En el ejemplo de la teoría de grupos se ve claramente que las disciplinas algebraicas individuales proporcionan no sólo los instru-

mentos para cálculos técnicos, sino también el lenguaje apropiado para la expresión de profundas leyes de la naturaleza. Sin embargo, aparte del valor práctico inmediato que numerosas ramas del álgebra encierran para la física, química, cristalografía y otras ciencias, el álgebra ocupa uno de los lugares más importantes en las propias matemáticas. Con palabras del conocido algebrista soviético N. G. Chebotarev, el álgebra ha sido la cuna de muchas de las ideas y conceptos que surgen en matemáticas y ha fertilizado en gran medida el desarrollo de ramas de las matemáticas que sirven de base directa a las ciencias físicas y tecnológicas.

INDICE ANALITICO

A

- Abel, N. H., I: 99, 318, 332, 333, 335; II: 171; III: 332, 343
Abscisa, I: 260
Adams, I., I: 21
Adición, III: 26
Aditiva, propiedad, II: 238
Aditividad total de la medida, III: 43
Ado, I. D., III: 396
Afin, clasificación de curvas, figuras, superficies, etc., I: 283
espacio, III: 77, 164
propiedades de las figuras, I: 283
Agregado; *véase* Conjunto
Al-Biruni, I: 60
Aleksandrov, A. D., I: 302; II: 140
Aleksandrov, P. S., III: 41, 195, n. 3 p. 239, 260, 261, 265, 266, 373
Alexander, J., III: 260, 261
Algebra, I: 63-4, 82-3, 315-6; III: 309, 385
abstracta, I: 319
asociativa, III: 385
axiomática, I: 319
base de, III: 387
con división, III: 394
estructura de, III: 393
topológica, III: 261
ideal de, III: 392
isomorfa, III: 388
de Lie, III: 395
lineal, III: 55
de matrices, III: 390-391
nihilpotente, III: 394
rango de, III: 387
semisimple, III: 394, 398
simple, III: 394
sistemas hipercomplejos, III: 310, 385
teorema fundamental del, I: 338-52
vectorial, I: 261; III: 381
Algebraico(a)
complemento, III: 87
entero, II: 240
operación, III: 406-407
solución, I: 320
variedad, III: 402
Algebrización, I: 316

- Algoritmo, II: 238, 239
 Eucúldeo, II: 238
 Al-Kharizmi, I: 61, 62, 63, 317
 Alternado, grupo, III: 327
 Análisis, I: 68, 73, 83, 91
 funcional, I: 84, 311-12; III: 265, 267
 Analítica, geometría, I: 68, 70, 225, 228-9
 en el espacio, I: 266-7
 infinito dimensional, I: 311
 prolongación, II: 222, 224, 225
 superficie, II: 125
 teoría de números, II: 240
 Analiticidad de una función compleja, II: 219
 Andronov, A. A., I: 396, 424; III: 254, 255
 Ángulo, entre vectores, III: 77, 270
 de paralelismo, III: 133-134, 136, 138
 Anillo, III: 398
 de clases residuales, III: 400
 factor, III: 400
 no conmutativo, III: 403
 de operadores, III: 404
 principales, III: 401
 semisimple, III: 404
 sin divisores de cero, III: 401
 Apolonio, I: 57, 58, 70
 Aproximación, por defecto, I: 364, 366
 por exceso, I: 364, 367
 fórmula de, II: 356
 de funciones, II: 311
 de integrales definidas, II: 324
 métodos de, II: 355
 óptima, II: 313, 337
 óptima uniforme, II: 313, 330, 340
 por polinomios, II: 314, 333
 Arco, longitud de, II: 82
 Argumentos de un número complejo, I: 340
 Aristóteles, I: 92
 Aritmética, I: 24
 Armónica, función, II: 30
 Arquímedes, I: 33, 57, 58, 59, 71, 91, 97, 167; III: 31
 principio de, III: 31
 Aryabhata, I: n. 18 p. 60
 Asintótica, convergencia, III: 135
 Asintótica, ley, II: 246
 Asintótica en un punto, II: 77
 Asociados, elementos, III: 401
 Asociativa, álgebra, III: 385
 Asociatividad, III: 26
 Autofunción, II: 47; III: 295
 Autointersección, III: 238
 Automorfismo, III: 317
 Automorfismo, grupo de, III: 345
 Autooscilaciones u oscilaciones propias, I: 391
 Autovalor(es)
 espectro continuo de, III: 304
 de operadores, III: 301
 de una transformación lineal, III: 105
 Autovector(es), III: 104-105
 de operadores, III: 301-302
 de una transformación lineal, III: 104-105
 Autovibraciones, I: 391; III: 293
 Axioma(s), de adición, III: 26
 de continuidad, I: 347; III: 29, 153
 de distributividad, III: 27
 de un espacio lineal, III: 68
 de geometría, III: 151
 de incidencia, III: 152
 del movimiento, III: 153
 de multiplicación, III: 26
 de orden, III: 27, 152
 de paralelismo, III: 153
 de teoría de la probabilidad, II: 272
 Axiomática, álgebra, I: 319

B

 Banach, S., I: 84
 Base, de un álgebra, III: 387
 de un espacio, III: 71-72
 Básico, intervalo, II: 257
 Bernoulli, D., I: 74, 98, 369; III: 281
 Bernstein, S. N., II: 290, 307, 314, 329, 336, 340
 Bernstein, polinomio de, II: 339, 340
 Bertrand, J. L. F., II: 245

Bessel, F. W., I: 222
 Bézout, E., I: 318
 Bhaskara, I: n. 18 p. 60
 Birkhoff, G., I: 425; III: 260
 Blaske, W., II: 143
 Bogoljubov, N. N., I: 424, 425; II: 302
 Boltjanski, V. G., III: n. 12 p. 260, 261
 Bolyai, F., III: 127, 131, 375
 Bolyai, J., I: 80; III: 131
 Bolzano, B., I: 83, 347, 348; III: 16, 36
 Bonnet, P. O., II: 128
 Borel, E., I: 83; II: 314
 Boyle, K., I: 105, 293; II: 29; III: 174
 Brahe, T., I: 242
 Brahmagupta, I: n. 18 p. 60
 Braquistócrona, II: 145, 146
 Briggs, H., I: 64
 Brouwer, L. E. J., III: 259, 260
 Budan, I: 357
 Bunjakovski, V. Ja., III: 79, 80, 276, 277

C

Cadena, III: 248
 bidimensional, III: 249
 unidimensional, III: 248
 Caja canónica, III: 108
 Cálculo, diferencial, I: 71, 166
 integral, I: 71, 72, 166
 matricial, III: 117
 técnicas de II: 355
 Cálculo de variaciones, II: 145
 funciones admisibles, II: 152
 problema variacional en, II: 157, 159
 soluciones aproximadas en, II: 164, 165
 Calor, vector conducción de, II: 21
 Cambio de variable, I: 175-6
 Campo(s)
 algebraico, III: 27
 continuo de direcciones, III: 252
 de descomposición, III: 344
 de direcciones, I: 398; III: 252-254
 escalar, II: 16
 vectorial, II: 15; III: 252
 continuo, III: 252
 Canónica, forma de una ecuación, I: 254
 de una matriz, III: 111-112
 forma cuadrática, III: 113
 matriz de Jordan, III: 108
 Cantor, G., I: n. 14 p. 47, 79, 83; III: 18, 29, 31-34, 37, 41, 259
 conjunto perfecto de, III: 39
 principio de continuidad de, III: 29
 Cantoriana, continuidad, III: 29
 de una figura, III: 194
 Caracteres, de grupos, III: 367
 Característica(o), ecuación, de una ecuación lineal homogénea, I: 389
 ecuación cuadrática, I: 292
 frecuencia; véase Autofunciones
 función; véase Autofunciones
 oscilaciones, I: 391
 Cardano, G., I: 318, 321, 322, 324, 325
 Cartan, E., III: 214, 358, 398
 Cauchy, A., I: 74, 83, 99, 117, 119, 171, 172; II: 180, 183, 187, 194, 198, 204; III: 16, 30, 31, 79, 80, 276, 277
 principio de, III: 30
 Cauchy-Riemann, ecuaciones de, II: 179, 206
 Cavalieri, B., I: 71
 Cayley, A., III: 169, 351
 Centro, I: 415
 Cero, I: 32
 Cerrado(a), clase de curvas o superficies, II: 139
 conjunto, III: 37, 262
 sistema ortogonal de funciones, III: 287
 superficie, III: 236, 244, 246
 superficie convexa, II: 140
 Cíclica, permutación, III: 328
 Cíclico, subgrupo, III: 331
 Ciclo(s), fundamental, III: 251-252
 límite, I: 421; III: 253
 límite estable, I: 422
 longitud de, III: 329-330
 permutación en, III: 328
 en topología, III: 250
 uni-dimensional, III: 249

- Ciclotómicas, ecuaciones, I: 332
 Cilíndricas, funciones, III: 282
 Cilindro elíptico, I: 270
 hiperbólico, I: 270
 parabólico, I: 270
 Circulación de flujo, II: 192
 Circulatorio, movimiento, II: 191
 Círculo de convergencia, II: 173
 Círculo límite, III: 139
 Circunferencia, no orientada, III: 337
 orientada, III: 336
 Clairaut, A. C., II: 178
 Clase cerrada de curvas y superficies, II: 139
 Clase, III: 352
 a la derecha, III: 352
 a la izquierda, III: 352
 Cociente por la derecha, III: 332
 Cociente por la izquierda, III: 332
 Cociente de diferencias de una función, II: 60
 Codazzi, II: 80, 128, 134
 ecuaciones de, II: 80, 128, 134
 Coeficiente de dilatación, II: 196
 363
 Coeficiente de una matriz, III: 62
 de transformación, III: 101
 Cohn-Vossen, S. E., III: n. 2 p. 343
 Cohomología, III: 261
 Colón, I: 91
 Columna, de una matriz, III: 59
 Combinatoria, topología, III: 260
 261
 Complejo(a), raíz, I: 327
 variable(s), ecuaciones de Cauchy-Riemann para, II: 179
 coeficiente de dilatación en, II: 196
 coordenadas, I: 312
 espacio, I: 312
 número, III: 374-375, II: 167-183
 argumento del, I: 340
 Complejo, plano, I: 352
 Complejos, enteros, II: 265, 267
 Completo, sistema de sucesos, II: 273
 Completitud de la recta numérica, III: 33
 composición de grupos, III: 347
 Computacional, matemática, I: 86
 Concavidad, I: 146-7
 Condición necesaria de convergencia, I: 209
 Conexión, III: 194
 Conexión de superficie, III: 235
 de variedades, III: 251
 Conforme, geometría, II: 143
 Conforme, transformación, II: 197, 200
 Congruencias, II: 137
 Congruencias de rectas, II: 137
 Congruentes, superficies, II: 128
 Cónicas, secciones, I: 252
 foco de, I: 248
 Conjunto(s), abierto, II: 37
 acotado, III: 33-34
 inferiormente, III: 34
 superiormente, III: 34
 acotado, e infinito de puntos, III: 36
 adherente, III: 194, 263
 aplicación unívoca de, III: 353
 cardinalidad de un, III: 22
 cero-dimensional, III: n. 11 p. 260
 cerrado, III: 37
 colineal de puntos, I: 298
 complemento de un, III: 38
 definición inductiva de dimensión de, III: n. 11 p. 260
 densidad de un, III: 27
 envoltura de, III: 42
 extremo inferior de, III: 34
 extremo superior de, III: 30-34
 finito, III: 20
 de funciones lineales, III: 57
 igual, III: 19
 infinito, III: 20, 36
 intersección de, III: 19
 lineal, III: 33
 linealmente ordenado, III: 404
 medible, III: 43
 n-dimensional, III: n. 11 p. 260
 no acotado, III: 33
 numerable, III: 23-24
 números racionales, III: 24
 ordenado, III: 33, 404
 parcialmente ordenado, III: 404
 perfecto, III: 39
 teoría de, III: 17

- teoría descriptiva de, III: 40
- totalmente ordenado, III: 404
- unión de, III: 19
- vacío, III: 19
- Conmutación, III: 396
- Conmutador, III: 340
 - de elementos de grupo, III: 333
- Conmutadores, subgrupo de, III: 333
- Conmutatividad, III: 26
- Conmutativo, grupo, III: 332
- Cono de segundo orden, I: 271
- Constante de estructura, III: 388
- Continuo(a), aplicación
 - función no regular, II: 323
 - funcional, III: 298
 - funciones, I: 117
 - grupo, III: 356, 358
 - grupo de transformaciones, III: 356
 - en un intervalo, I: 118
 - lateralmente, I: 119
 - en un punto, I: 118
 - representación, III: 358
 - transformación, III: 195, 232, 233, 263; II: 126
 - de medio continuos, I: 284-285; III: 257
 - variable, I: 54
 - no numerabilidad del, III: 32
 - número cardinal del, III: 25
- Contorno, condiciones de, II: 31, 34
 - condiciones homogéneas de, II: 36
 - problema de, II: 29, 33, 155
- Contracción del plano a una línea, I: 276, 285
- Contracción, teoremas de, II: 364
- Convergencia, asintótica, III: 135
 - círculo de, II: 173
 - dominio de, II: 171
 - «en media», III: 283
 - rapidez de, II: 361
 - de series, I: 209
- Convergente, onda esférica, II: 43
- Convergentes, métodos, II: 361
- Convergentes, series, I: 209
 - absolutamente, I: 211
- Convexa hacia abajo, I: 147
- Convexidad, I: 146-147
- Convexo, cuerpo, II: 140-142
- Coordenadas, curvilíneas, II: 114, 129
 - método de, I: 229
 - polares, I: 310
 - de un punto, I: 227
 - rectangulares, I: 260
 - sistema de, I: 281
 - transformación de, I: 255; III: 102-103
 - vectoriales, I: 263; III: 74
- Copérnico, N., I: 91, 225
- Correlación, coeficiente de, II: 288
- Correspondencia, uno a uno, III: 21
- Corriente, línea de, II: 185
- Cortadura de Dedekind, III: 29
- Corte en funciones complejas, II: 182
- Cota, I: 260
- Cota superior, III: 30
- Coulomb, C. A., I: 76
- Criba de Eratóstenes, I: 58; II: 241
- Cuadrática, ecuación, I: 321
 - característica, I: 292
- Cuaterno, norma de, III: 380
- Cuaternos, parte vectorial de, III: 380, 382-383
 - parte escalar de, III: 380
- Cuaternos conjugados, III: 378-379, 380-381
- Cúbica, ecuación, I: 321
- Cubo tetradimensional, III: 177
- Cuerpo, III: 378
 - algebraico, III: 27
 - ordenado, III: 27
 - de base de álgebra, III: 344, 386
 - no simétrico, III: 319
 - simétrico, III: 318
- Curva(s), clases afines de, I: 283
 - clases cerradas de, II: 138
 - curvatura de, II: 87-90
 - ecuación de, I: 266
 - equidistante, III: 135
 - en el espacio, longitud del arco de, II: 83
 - familias de, II: 234
 - integral, I: 399; III, 253
 - irreducible, III: 403
 - longitud de, II: 83

matemática, II: 75
 n -dimensional, III: 184
 de nivel, I: 181, 267
 Curvatura, cero de una curva, II: 87, 90
 constante, III: 211
 del espacio, III: 209
 Gaussiana, II: 108
 media, II: 108
 principal, II: 103
 sobre una superficie, II: 101-108
 total; véase Curvatura Gaussiana
 Curvatura, tensor, III: 208
 Curvilínea(s), coordenadas, II: 114
 129
 integral, I: 201-203

CH

Chaplygin, S. A., II: 191, 193
 Chebotarev, H. G., I: 337, 338;
 III: 396, 402, 408
 Chebichev, P. L., I: 84, 87, 99,
 178; II: 79, 240, 243-246, 249,
 252, 263, 286-290, 329-334,
 335, 341, 350, 351; III: 282
 aproximación de, II: 313, 330-
 332, 340
 desigualdad de, II: 287
 polinomios de, II: 322, 323, 334
 Chech, E., II: 143
 Chudakov, N. G., II: 247

D

D'Alembert, J., I: 211; III: 127,
 281
 criterio de series convergentes,
 I: 211
 lema de, I: 349
 Dandelin, G. P., I: n. 14 p. 369
 Darboux, G., II: 80, 125
 Darwin, C. R., III: 131
 Da Vinci, L., I: 91; III: 160
 Dedekind, R., I: n. 14 p. 47, 83;
 II: 240; III: 29-32, 406
 Dedekind, continuidad de, I: 347;
 III: 29-30, 153
 cortadura, III: 28-29

 principio de, III: 29
 retículo, III: 406
 Definida, integral, I: 165
 Deformabilidad de una superficie,
 II: 142
 Deformación de una superficie, II:
 115, 125, 142
 Deformación uniforme de una
 membrana, II: 147
 del Ferro, Scipio, I: 321, 329
 de Jonquières, III: 245
 Demócrito, I: 40, 45, 53, 54
 Densidad lineal, III: 292
 Dependencia, funcional, II: 149-
 lineal, III: 69
 Derivación, reglas de, I: 132
 término a término, I: 212
 Derivada, de un cociente, I: 134
 de una función, I: 124
 de orden superior, I: 145
 parcial, I: 182
 parcial cruzada, I: 184
 de un producto, I: 133
 segunda, I: 146
 de una suma, I: 132
 Desargues G., III: 160
 Desarrollo,
 de una función en serie trigono-
 métrica, II: 243; III: 284
 en un sistema ortogonal de fun-
 ciones, III: 279, 282
 Descartes, R., I: n. 13 p. 46, 59,
 63, 68, 69, 70, 85, 96, 226-231,
 233-234, 236, 237, 260, 316, 317,
 318, 355-358; II: 129; III: 243
 Descartes, ley de los signos de, I:
 355
 Descomposición, de números ente-
 ros, II: 265
 de una permutación en ciclos,
 III: 328
 simplicial, III: 247
 de una variedad, III: 403
 Desviación
 típica, II: 287
 de frecuencia, II: 286
 probable, II: 276
 Determinados, procesos, II: 299
 Determinante(s), de los coeficien-
 tes de una matriz, II: 89

- complemento algebraico en, III: 87
- desarrollo de, III: 89
- de una matriz, III: 83
- menor, III: 87
- de orden n , III: 86-87
- Diagonal, proceso, III: 24
- Diámetro, de una región, I: n. 34 p. 200
- teoría de Newton de, I: 237
- Diferencia de dos conjuntos, III: 20
- Diferenciabilidad de una función, II: 177
- Diferenciable, función, I: 129; II: 195
- Diferencial(es), ecuación(es), I: 373 y sigs.
- campo de direcciones, I: 398; III: 253-254
- centro de, I: 415; III: 258
- en derivadas parciales, I: 376
- estable, I: 417
- estocástica, II: 307-309
- de Euler, II: 151-154
- foco en, I: 415
- lineal, I: 386-387
- lineal homogénea, I: 387
- orden de, I: 384
- ordinaria, I: 373, 376, 383
- «poco sensibles», I: 396
- solución general de, I: 385
- Diferencial de una función, I: 150
- de varias variables, I: 185
- Dilatación, coeficiente de, II: 196, 363
- Dimensión(es), del espacio, III: 71
- de una variedad, III: 242-243
- Dimensión, teoría de, III: 259
- Diofanto, I: 59, 60
- Dirección, números de, I: 407
- principal, II: 103
- Directos, métodos, I: 402
- Directriz de una parábola, I: 248
- Dirichlet, P. G. L., I: 108; II: 74; III: 46, 49
- función de, III: 46
- Discontinua, función, I: 119
- Discontinuidad débil, II: 68
- Discreto, grupo de movimientos, III: 338, 342
- Discreto, grupo simétrico, III: 338
- Disjuntas, clases, en grupos, III: 349
- Disjuntos, conjuntos, III: 20
- Distancia, en un espacio n -dimensional, III: 262
- entre funciones, III: 263-264
- entre puntos, III: 33
- Distribución de números primos, II: 242-244
- Distributiva, ley, III: 20
- Distributividad, III: 27
- Divergencia, teorema de Gauss de; véase Ostogradski, fórmula de
- Divergentes, series, I: 207
- Divisibilidad, II: 236
- División, álgebra con, III: 394
- Divisor, de cero, III: 401
- normal, III: 352
- de uno, III: 401-402
- Dominio, de convergencia, II: 171
- de definición, II: 149; III: 47, 299
- de una función, I: 103
- integral, III: 401
- de racionalidad, III: 344
- simplemente conexo, II: 201
- de valores, III: 191
- de variación de variables, III: 191
- Dualidad, teorema de, III: 261
- Durero, A., III: 160

E

- Ecuación(es), característica, homogénea lineal, I: 389
- de Cauchy-Riemann para variables complejas, II: 179, 206
- centro en, I: 415
- ciclotómica, I: 333
- de conservación de la energía calorífica, II: 22
- de continuidad, II: 20
- cuadrática, I: 321
- de cuarto grado, I: 324
- de una cuerda vibrante, II: 26
- de una curva, I: 266
- diferencial, I: 373-425

- construcción de soluciones, II: 43
- estocástica, II: 307-309
- estructuralmente estable, I: 417
- lineal, I: 386
- lineal homogénea, I: 387
- lineal no homogénea, I: 387, 392-395
- ordinaria, I: 373, 375, 384
- solución general de, I: 385
- forma canónica de, I: 255
- funcional, I: 374
- homogénea correspondiente, II: 43
- integral, III: 288-296
 - homogénea, III: 292
 - no homogénea, III: 294
 - de segunda especie, II: 57
- de Laplace, II: 30, 194
 - de una línea recta, I: 268
 - del movimiento, II: 23
 - de oscilación, II: 35
- parcial derivada, I: 376; II: 15
 - soluciones admisibles o generalizadas de, II: 65, 69
 - construcción de soluciones, II: 43
- paramétrica, II: 81
- del plano, I: 268
- de Poisson, II: 30
- de primer grado, I: 231, 320
- de Ricatti, I: 395
- secular, III: 106
- de segundo grado, I: 320
- solución algebraica de, I: 320
 - de superficie, I: 266
 - de tercer grado, I: 321
- Efimov, N. V., II: 127
- Egorov, D. F., III: 47
- Einstein, A., I: 301; III: 131, 200, 216, 223-227
- Eje, de grupo, III: 338-339
 - de rotación, III: 312
 - de simetría, III: 312
- Elemento(s), asociado, de un anillo, III: 401
 - cero, III: 348
 - conjugados, de un grupo, III: 349
 - de un conjunto, III: 18
 - de longitud, III: 205
 - de una matriz, III: 58
 - indescomponible, III: 401
 - inverso, III: 99
 - neutro, III: 348
 - primo, III: 401
 - de superficie, II: 120, 132
 - unidad, III: 348-349
- Elipse, I: 240-241
 - imaginaria, I: 254
 - de inercia, I: 243-245
- Elipsoide, I: 269
 - imaginario, I: 269
- Envolvente de una familia de curvas, II: 135
- Equivalentes, caminos, III: 360
- Eratóstenes, I: 58; II: 241
 - criba de, I: 58; II: 241
- Escalar, I: 262; II: 16
 - multiplicación, III: 80
- Escalar, producto, espacio n -dimensional, III: 271
 - de vectores, I: 262; III: 78, 383
- Espacio, abstracto, III: 164
 - transformación de, I: 280-281; III: 162
 - base de, III: 71-72
 - configuración de, III: 213
 - de color, III: 185
 - curvatura de, III: 209-210
 - de curvatura cero, III: 211
 - curvatura constante, III: 211
 - dimensión de, III: 71, 259
 - de dimensión infinita, III: 71, 191, 196, 273-279, 298
- Euclidiano, III: 80
 - fásicos, I: 410; III: 187, 243
 - funcional, III: 192, 265, 299
 - de infinitas dimensiones, III: 299
 - generalizado, III: 190
 - de Hilbert, III: 196, 273-379
 - lineal complejo, III: 76
 - lineal general, III: 68
 - lineal de dimensión infinita, III: 298
 - lineal de dimensión finita, III: 386
 - métrico, III: 196, 262-263
 - multidimensionales, III: 168, 174
 - n -dimensional, III: 175

- real, III: 215
- de Riemann, III: 199-202
- rotación de, III: 318
- solución, III: 97
- topológico, III: 194, 262, 265
- topológico abstracto, III: 164
- unitario, III: 82
- vectorial lineal, III: 68-69
- Espacio-tiempo, III: 222
- Espectro continuo, III: 304-305
- Esperanza matemática, II: 287
- Estimación de un error, II: 360
- Estabilidad de los procesos de cálculo, II: 369
- Euclides, I: 22, 34, 40, 41, 46, n. 13 p. 46, 57, 58, 63, 77; II: 241, 244, 247; III: 123-130, 140, 141, n. 7 p. 144, n. 10 p. 151, 152, 154, 160, 164, 211, 215, n. 37 p. 215, 217, 220
- Euclides, postulado de las paralelas de, III: 124, 153
- Eudemo de Rodas, I: 39
- Eudoxio, I: n. 12 p. 46
- Euler, L., I: 75, 78, 98, 127, 201, 253, 259, 260, 317, 318, 402, 403, 407; II: 54, 78, 103, 105, 107, 114, 125, 134, 151, 154-157, 159, 161, 175, 176, 240, 242, 244, 246, 249, 260, 315, 343, 370; III: 243-247, 252, 256 281
- Euler, ángulos de, I: 269
- características de triangulación, III: 244
- ecuación diferencial de, II: 151, 154
- fórmula(s) de, II: 175; III: 244
- función(es) de, II: 260
- método de las líneas quebradas de, I: 403
- teorema de, II: 103, 105
- Euler-Poincaré, fórmula de, III: 252
- Excentricidad de la sección cónica, I: 251
- Existencia de una solución, I: 401
- Extremas, condiciones, II: 150
- Extremos, valores de funcionales, II: 149
- Extrínseca, propiedad de una superficie, II: 127
- F
- Faber, G., II: 323
- Factor irreducible, I: 334
- Factorización de un polinomio, I: 326
- Falsa, posición, I: 366, 368
- Familia de curvas, II: 134
- n -paramétrica, II: 134
- de superficies, II: 134
- Fedorov, E. S., I: 42, 82, 87; III: n. 23 p. 180, 311, 340, 342, n. 2 p. 343
- Fejér, L., II: 349, 352
- Fejér, suma de, II: 349
- Fermat, P., I: 226, 316, 317; II: 265; III: 402
- Ferrari, I: 63, 324
- Figuras, clasificación afín de, I: 283
- planas infinitas, III: 337
- propiedades afines de, I: 283
- Finikov, S. P., II: 143
- Finsler, R., I: 81; III: 214
- Flujo, II: 20
- laminar, II: 188
- rotacional, II: 26
- vector de, II: 20
- Foco, en ecuaciones diferenciales, I: 415; III: 258
- de una sección cónica, I: 247
- Fok, V. A., III: 226
- Fokker, II: 307
- Forma cuadrática, III: 109, 114
- canónica, III: 113
- equivalente, III: 111
- fundamental, III: 204
- inercia de una, III: 113
- primera fundamental, II: 131-134
- segunda fundamental, II: 133
- Forma lineal, III: 58
- Fórmula, de los incrementos finitos, I: 154
- fundamental del cálculo integral, I: 205
- Fourier, J. B. J., I: 358; II: 44, 54, 58, 315, 339, 341, 344, 347-349; III: 53, 196, 304
- Fourier, coeficientes de, II: 344
- integral de, III: 304
- series de, II: 341-349

- Fracciones, origen de, I: 43
 Frecuencia característica, *véase* Autofunción
 Frecuencias, espectro de, III: 293
 Frecuencia propia, III: 293
 fundamental, II: 51
 Frenet, J. F., II: 79
 Fridman, A. A., III: 227
 Frontera, de una cadena, III: 249
 de un conjunto, III: 194; *véase también* Punto frontera
 de un triángulo orientado, III: 247-248
 Fubini, G., II: 143
 Función(es), admisible, II: 152
 analítica, I: 219; II: 170
 completa, II: 224
 en intervalos, I: 220
 en un punto, II: 177
 representación de, II: 221
 aproximación de, II: 311
 aproximadora, II: 331
 armónica, II: 30
 máximo o mínimo absoluto de, II: 32
 característica, II: 47; III: 49
 de flujo, II: 187
 cociente de diferencia de, II: 60
 compleja, II: 167-231
 compuesta, II: 130
 constante a trozos, II: 68
 continua, I: 117
 a un lado, I: 119
 no regular, II: 139, 323
 continuamente diferenciable, II: 71; III: 55; *véase también* función continua
 de corriente, II: 185
 de cuadrado integrable, III: 299
 definición implícita de una, I: 181
 derivada de, I: 121
 derivable, I: 129; II: 195
 diferencial de, I: 150
 de Dirichlet, III: 46
 discontinua, I: 119
 dominio de, I: 103
 elemental, I: 139; II: 169
 entera, II: 216
 esférica, III: 281
 de Euler, II: 260
 exponencial, II: 174
 de matrices, III: 118
 fraccionaria, II: 218
 de una función, I: 137
 gráfica de, I: 103, 140-9
 de Green, III: 288-292
 impar, II: 237
 implícita, definición de, I: 181
 implícita, derivación de, I: 187
 integrable, I: 172, 196
 integral, II: 215
 inversa, I: 135
 de Lagrange, II: 70
 límite, III: 16
 lineal, I: 102; III: 55
 lineal homogénea, III: 56
 lisa, II: 68; III: 55, 98
 lisa a trozos, II: 68
 de una matriz, III: 117
 argumento, III: 117
 medible, III: 45
 multiforme, II: 181, 225
 no regular, II: 139
 ortogonales, III: 278
 ortogonales dos a dos, III: 279
 par, II: 237
 prolongación de, en un dominio complejo, II: 170
 regular, II: 322
 sistema ortogonal cerrado de, III: 287
 sistema ortogonal infinito de, III: 287
 de teoría de números, II: 260
 teoría trascendente de, II: 195
 trascendente, II: 217
 trigonométrica, II: 174
 de una variable compleja, II: 174
 de variable real, III: 15
 de varias variables, I: 179
 vectorial, II: 130
 Funcional, I: 155; III: 297

G

- Galerkin, B. G., II: 58, 59, 73, 358
 Galerkin, método de, II: 58
 Galileo, I: 65, 68, 70, 92, 93, 225, 301; II: 300, 301

- Galois, E., I: 82, 318, 330-8; III: 343-347
- Galois, grupo de, I: 355; III: 343-345
- teoría de, I: 332-6
- Gauss, K. F., I: 98, 318, 332, 336; II: 111-114, 128, 130, 134, 240, 244-330; III: n. 2 p. 126, 127, 131, 199, 209, 210, 216
- curvatura de, II: 108, 111; III: 209-210
- teorema de la divergencia de; véase Ostrogradski, fórmula de
- Gelfand, I. M., III: 358
- Gelfond, A. O., II: 240
- Generadores, de grupos, III: 363
- libres, III: 364
- de subgrupo, III: 330
- Generalizadas, soluciones de ecuaciones diferenciales parciales, II: 67, 69
- Generadores, de grupo, III: 363
- Generatrices rectilíneas, I: 274
- Género de una superficie, II: 230; III: 236
- Geodésica, II: 122; III: 143, 206
- Geodésica(o), curvatura, II: 119
- línea, II: 121
- local, II: 137
- superficie, III: 210
- segmento, II: 119
- Geometría, I: 38, 80
- abstracta, III: 215
- afín, III: 162
- algebraica, I: 311
- analítica, I: 68, 70, 225, 228, 311
- en el espacio, I: 266-7
- analítica de infinitas dimensiones, I: 311
- axiomas de la, III: 151
- conforme, II: 143
- diferencial, II: 77, 80-82, 143
- afín, II: 143
- clásica, II: 137
- proyectiva, II: 143
- euclidiana, III: 141
- global, II: 137
- intrínseca, II: 114-134; III: 201
- de Lobachevski, III: 132
- n-dimensional, III: 169
- no euclidiana, III: 123
- proyectiva, I: 293, 299; III: 159
- riemanniana, III: 199
- Geométrico(a), figura, I: 38
- interpretación, I: 398-400
- Gibbs, J. W., III: 174
- Gjunter, N. M., II: 72
- «Global», I: 420; II: 77, 136
- Goldbach, C., II: 249, 256, 259, 264
- Goldbach, conjetura de, II: 249, 256
- Gradiente de temperatura, II: 28
- Grado, de una representación mediante matrices, III: 367-392
- de simetría, III: 319
- Graduada, representación, III: 370
- Graeffe, I: n. 14 p. 369
- Graeffe, método de, I: 369
- Gráfica, de una función, I: 103, 104, 140-149
- de un polinomio, I: 364
- Grassmann, H., II: 129; III: 169
- Green, G., III: 288
- Green, función de, III: 288-292
- teorema de; véase también Ostrogradski, fórmula de
- Grupo(s),
- abeliano, III: 332
- abstracto, III: 373
- alternado, III: 327
- de automorfismos, III: 345
- axiomas de, III: 348
- de caminos, III: 361
- carácter de, III: 367
- clases disjuntas en, III: 349
- composición de, III: 347
- conmutativo, III: 332
- caracteres de, III: 261
- continuo, III: 356, 358
- cristalográfico, III: 334, 340, 342
- definición, general, III: 347
- derivado, III: 333
- de ecuaciones algebraicas, III: 344
- eje de, III: 338-339
- elemento conjugado en, III: 349
- factor, III: 351
- de Fedorov, III: 334, 340, 342
- finito, III: 373
- finito simple, III: 358
- fundamental, III: 359

- de Galois, I: 355; III: 343-344
 generadores de, III: 363
 infinito, III: 373
 isomorfo, III: 349
 libre, III: 364
 de Lie, III: 356, 358
 local, III: 396
 de movimientos, III: 338
 multiplicación de, III: 347
 de nudos, III: 365, 366
 operación de, III: 348
 orden de, III: 325
 de permutaciones, III: 327
 plano de Fedorov, III: 334, 340, 341, 342
 relaciones de definición de, III: 363
 representación de, III: 367
 representación graduada en, III: 370
 representación por matrices, III: 367
 de rotaciones, III: 326
 simétrico, de permutaciones, III: 327, 347
 suma de, III: 348
 simple, III: 352
 topológico, III: 358
 de transformaciones, III: 321-334, 347
 en geometría diferencial, II: 143
 Gyaseddin Jamschid, I: n. 18 p. 60, 62
- H
- Hadamard, J., II: 246
 Hamilton, W. R., II: 70-72, 130, 160
 Hamilton, principio de; véase Principio de Ostrogradski-Hamilton
 Hardy, G. H., II: 240
 Hermitiano, espacio vectorial; véase Espacio unitario
 Hertz, H., I: 21
 Hilbert, D., I: 84, 99, 337, 338
 espacio de, III: 196, 273-279
 Hincin, A. Ja., I: 425; II: 289
 Hiparco, I: 57
 Hipérbola, I: 247
 rectangular, I: 253
 Hiperboloide, I: 269, 273
 de revolución, I: 273
 Hipercomplejos, números, III: 374-385
 sistemas, III: 310, 385
 Hiperplano, III: 171
 tangente, III: 184
 Hipócrates, I: 56
 Holónomo, sistema, II: 160
 Homeomorfas, figuras, III: 235
 Homología, III: 250
 Homológica, independencia de los ciclos, III: 250
 teoría de la dimensión, n. 11 p. 260
 Homólogo a cero, III: 250
 Homomorfa, aplicación, III: 353
 imagen, III: 256
 Homomorfismo, III: 353
 Homotópica, topología, III: 261
 Hooke, R., III: 290
 Hopf, H., III: 256, 260, 261
 Hopf, teorema de, III: 256
- I
- Ideal, III: 399
 bilátero, III: 393
 cero, III: 393
 de elementos de un álgebra III: 392
 a derechas, III: 392
 a izquierdas, III: 393
 minimal, III: 393
 principal, III: 399
 indescomponible, III: 403
 Ideales, líneas, I: 297
 puntos, I: 298
 Imagen especular, III: 311
 Incomensurables, magnitudes, I: 44
 Incremento, I: 118, 150
 Independencia, III: 57
 lineal, III: 69
 Independientes, sucesos, II: 275
 Índice, de un punto singular, III: 256

de un punto singular aislado, III: 256
 Infinitamente grande, I: 112
 Infinitésimo, I: 110
 de orden superior, I: 151
 Inflexión, punto de, I: 148
 Inicial, condición, II: 34
 estado, I: 402; II: 29
 valor de un problema, I: 402; II: 29
 Integración, de caminos, II: 210
 por partes, I: 178
 Integral, I: 163
 de acción, II: 161
 de Cauchy, II: 211
 a lo largo de una curva, II: 208
 curvilínea, I: 201-3
 definida, I: 165
 doble, I: 200
 indefinida, I: 173; II: 210
 de Lebesgue, III: 47-53
 de línea, I: 201; II: 207
 sobre un contorno cerrado, II: 211
 múltiple, I: 197
 de Riemann, I: 165; III: 47
 de superficie, I: 201
 Interpolación, lineal, I: 366, 368
 polinomios de, II: 316, 335
 Intersección, de conjuntos, III: 19-20
 de sucesos, II: 274
 Intervalo, III: 33
 abierto, III: 33
 adyacente, III: 38
 básico, II: 259
 cerrado, III: 33
 encajado, III: 29
 impropio, III: 33
 Invariante, I: 288
 topológico, III: 362
 Inverso, de un cuaternio, III: 381
 de una matriz, III: 93
 Irreducibilidad de un polinomio, I: 334
 Irrotacional, movimiento en un fluido, II: 187
 Isomorfa(s), aplicación, III: 350
 Isomorfismo, III: 349
 local, III: 396
 Iteración, método de, II: 362

J

Jackson, D., II: 314, 340
 Jordan, C., III: 108, 121
 Jukovski, N. E., I: 22

K

Keldys, L. V., III: 41
 Keldysch, M. V., II: 340; III: 305
 Kepler, J., I: 70, 71, 91, 96, 226, 242, 243, 375
 Killing, III: 358, 398
 Kirchhoff, G. R., II: 389
 Kiselev, A. P., I: 40
 Klein, F., I: 99, 338; III: 144, 150-154, 163, 200, 214, 225, 238, 257, 258, 358
 Klein, botella de, III: 256-257
 Kolmogorov, A. N., II: 307, 314, 315, 348; III: n. 1 p. 25, 41, 261
 Kolosov, G. V., II: 195
 Korkin, A. N., II: 314
 Krasnoselski, M. A., III: 265
 Krylov, A. N., I: 372, 405
 Krylov, N. M., I: 424, 425
 Kuiper, N. H., II: 142; III: n. 8 p. 144
 Kummer, E. E., II: 240; III: 402
 Kurnakov, N. S., III: 174
 Kurosch, A. G., I: 320; III: 373

L

Lagrange, J. L., I: 74, 78, 154, 193, 194, 195, 196, 260, 261, 318, 328-32, 334, 335, 337; II: 70, 240, 320, 329; III: 16, 127, 169, 223, 343, 352
 Lagrange, fórmula de, I: 154
 fórmula de interpolación de, II: 320
 función de, II: 70
 multiplicadores de, I: 78, 193
 resolventes de, I: 330
 teorema de, III: 352
 Laguerre, E., I: 260
 Lambert, J. A., III: 125, 126

- Laminar, flujo, II: 188
 movimiento, II: 183
 Laplace, P. S., II: 30-32, 40, 43, 52, 54, 62, 68, 194, 284, 303, 359, 382, 383
 ecuación de, II: 30
 Laplaciana, II: 30
 Lappo-Danilevski, I. A., III: 121
 Laptev, G. F., II: 143
 Lavrentiev, M. A., I: 404; II: 207, 314, 340; III: 41
 Lebesgue, H. I.: 83, n. 25 p. 169, 197; II: 72, 314, 348, 353; III: 17 41, 47-53, 279, 287
 suma integral de, III: 52
 Lefschetz, S., III: 260
 Legendre, A. M., II: 244, 330; III: 125, 126, 282
 polinomios de, II: 330; III: 282
 Leibnitz, G. W., I: 68, 71, 96, 97, 127, 167, 169, 170, 171, n. 27 p. 171, 173, 316; II: 78, 135; III: 131
 Leverrier, U. J. J., I: 21, 375
 Ley(es) de inercia, III: 113
 de los grandes números, II: 280, 289
 de probabilidad, II: 269
 triangular, III: 264
 L'Hôpital, II: 78
 Lie, S., I: 320; III: 261, 356-358, 389, 395-398
 Lie, álgebra de, III: 395
 grupo de, III: 356-358
 representación continua de, III: 358
 Límite, I: 72, 108-17
 de una sucesión de matrices, III: 117
 uniforme, II: 69
 Línea, ecuaciones de, I: 268
 infinitamente distante, III: 239
 en el infinito, III: 239
 impropia, III: 239
 integral, I: 399
 Linnik, Ju. V., II: 247; III: 385
 Lisma, II: 142
 Liouville, J., I: 333, 395
 Littlewood, J. E., II: 240
 Liapunov, A. M., I: 74, 84, 99, 416-9, 424; II: 58, 290; III: 41
 estable en el sentido de, I: 417
 no estable en el sentido de, I: 417
 Ljusternik, L. A., II: 138; III: 260, 262, 265
 Lobachevski, N. I., I: 22, 40, 80, 81, 99, 108, 310, 318, 369, 371, 372; III: 123, 127-159, 164, n. 28 p. 193, 211, 215, 216, 217, 220, 221, 222, 227
 geometría de, III: 132
 curva equidistante en la, III: 135
 método de, I: 369
 postulado de las paralelas de, III: 154
 «Locales», II: 77, 136
 Lógica matemática, I: 86
 Lomonosov, M. V., III: 185
 Lorentz, H. A., I: 276, 301, 304-10; III: 225
 transformación de, I: 301-3
 fórmulas de transformación de, I: 304
 Luzin, N. N., I: 83; II: 315; III: 41, 47

M

- Mach, E., III: 222
 Malsev, A. I., III: 358
 Mandelstam, L. I., I: 424; III: 254
 Máquina integradora, II: 381
 clasificadora, II: 378
 digital, II: 378
 Máquinas analógicas o de acción continua, II: 380
 Mariotte, E., I: 105, 293; II: 29; III: 174
 Markov, A. A., I: 99, 425; II: 289, 290, 306, 314, 336
 Markov, V. A., II: 314
 Markov, proceso de, II: 306
 Markusevic, A. I., III: n. 9 p. 259
 Matemática moderna, I: 87
 Matricial, álgebra, III: 389
 cálculo, III: 117
 elemento, III: 58
 notación, III: 92
 operación, III: 59

- representación, de álgebras asociativas, de grupos, III: 367
- Matriz (matrices), III: 58
 - antisimétrica, III: 397
 - canónica de Jordan, III: 108
 - de coeficientes, III: 62
 - determinante de, III: 90
 - columna de, III: 59
 - cuadrada, III: 59
 - potencias de, III: 117
 - determinante de, III: 83
 - diagonal, III: 108
 - ecuación característica de, III: 106
 - de igual estructura, III: 62
 - fila de, III: 59
 - función, III: 117
 - exponencial, III: 119
 - multiplicación de, III: 63
 - no degenerada, III: 94
 - no nula, III: 401
 - no singular, III: 94
 - nula, III: 63, 64
 - números característicos de, III: 106
 - ortogonal, III: 397
 - polinomio característico de, III: 106
 - rango de, III: 95
 - semejante, III: 104
 - de transformación de coordenadas, III: 103
 - traspuesta, III: 64
 - unidad, III: 93, 391
- Máximo, I: 140
- Máximo, valor de una función, I: 141
- Maxwell, J. C., I: 21; III: 187
- Medible, conjunto, III: 43
 - función, III: 45
- Medida de un conjunto, III: 41, 43
 - de un conjunto abierto, III: 42
- Medida exterior, III: 43
- Mendeleiev, D. I., II: 336
- Menshov, D. E., II: 315
- Mergeljan, S. N., II: 340
- Meromorfa, función, II: 219
- Método, de aproximaciones sucesivas, I: 405; II: 362
 - combinatorio en topología, III: 243-244
- de coordenadas, I: 229
- del «corto» y «largo», I: 366
- de las cuerdas, I: 366, 368
- de diferencias finitas o de redes, II: 58, 60
- geométrico, III: 197
- de las líneas quebradas (véase de las poligonales)
- de Newton, I: 368
- numéricos, II: 355
- de las poligonales, I: 403; II: 165, 166
- de las potenciales, II: 54
- de redes, II: 58, 60
- de sustitución, I: 175
- de las tangentes, I: 366
- Métrica, III: 201
 - intrínseca, III: 196
- Meusnier, J. B. M., II: 104, 105, 107, 108, 114, 134
 - teorema de, II: 104
- Michelson, A. A., I: 301, 302
- Miguel Angel, I: 91
- Minding, F., II: 79, 125, 137; III: 143
- Minimal(es), ideal, III: 393
 - principios, II: 130
 - problema, II: 147
 - superficie, II: 111
- Minimizadora, sucesión, II: 165
- Mínimo, I: 140
 - valor de una función, 141
 - problema del, para una integral múltiple, II: 161
- Minkowski, H., II: 139, 240; III: 223
- Möbius, A. F., I: 298; III: 237-240
 - banda de, III: 236-237
 - cinta de, III: 237
 - red de, I: 298, 299
 - superficie de, III: 236-238
- Modelo circular del plano proyectivo, III: 240-241
- Modular, retículo, III: 406
- Módulo, de un número complejo, I: 340
 - de un polinomio, I: 345
 - superficie de, I: 345, 351
- Moirve, de, II: 284
- Molin, F. E., III: 394, 395, 404

Monge, G., II: 137
 Monotonía, ley de, III: 27
 Moreland, II: 376
 Morozov, V. V., III: 358
 Morse, M., II: 138; III: 265
 Movimiento, ecuaciones de, II: 23
 helicoidal, III: 318
 impropio, III: 318, 342
 permanente, II: 183
 no perturbado, I: 418
 perturbado, I: 418
 plano-paralelo, II: 183
 de primera especie, III: 318
 propio, III: 318
 de segunda especie, III: 319
 Multiplicación, tabla de, III: 350
 Multiplicador, II: 378
 Multiplicativa, propiedad, II: 238
 Multiplicidad de una raíz, I: 353
 Mushelishvili, N. I., II: 195

N

Naimark, M. A., III: 358
n-dimensional, ángulo, III: 170
 conjunto, III: n. 11 p. 260
 cubo, III: 176
 cuerpo, III: 181
 curva, III: 184
 espacio, III: 268-269
 espacio euclidiano, III: 77, 80
 geometría, III: 170
 plano, III: 170
 punto, III: 66
 poliedro, III: 175-181
 segmento, III: 170
 simplex, III: 178, 247
 topología, III: 259
 triangulación, III: 247
 vector, III: 66, 270
 Nash, J., II: 142
 Nasireddin Tusi, I: 61; III: 125
 Nemyski, V. V., I: 425; III: 255, 265
 Neper, J., I: 64, n. 21 p. 64, 68
 Newton, I., I: 54, 61, 62, 64, 68, 71, 73, 74, 75, 78, 96, 97, 116, 167-73, 178, 228, 237, 243, 260, 316, 317, 318, 358, 368, 369, 375,

378; II: 23, 25, 29, 78, 176, 300, 313, 320; III: 131
 método de, I: 368
 potencial de, II: 55
 teoría de los diámetros de, I: 237
 Newton-Leibniz, fórmula de, I: 169
 No enumerabilidad del continuo, III: 32
 No orientabilidad, III: 246
 Nodo(s), I: 414; III: 253
 equivalentes, III: 365
 no equivalentes, III: 365
 de una oscilación, II: 51
 Normal, II: 99
 principal, II: 95
 Novikov, P. S., III: 41, 367
 Núcleo, de una ecuación integral, III: 294
 simétrico, III: 295
 Nudos, III: 365
 equivalentes, III: 365
 no equivalentes, III: 365
 Numerabilidad, III: 27
 Número(s), I: 27
 algebraico, II: 240
 analítico, II: 340
 de Betti, bidimensional, III: 251
 característico, III: 106
 cardinal del continuo, III: 25
 compuesto, II: 242
 de dirección, I: 407
 entero, I: 24; II: 235
 negativos, I: 55
 primo, II: 235, 237
 real, axiomas de continuidad de, I: 347; III: 29, 153
 teoría analítica de, II: 340
 elemental de, II: 339
 geométrica de, II: 339
 de vueltas del campo, III: 256

O

Odner, V. T., II: 377
 Omar Khayyam, I: 61, 62
 Onda esférica divergente, II: 43
 Operación(es), sobre conjuntos, III: 19

- sobre determinantes, III: 86
- m-aria, III: 407
- no conmutativa, III: 321
- sobre vectores, III: 65
- Operador(es), III: 299
 - autoadjunto, III: 297, 302, 303
 - lineal, III: 303
 - lineal, III: 296, 299
- Orden, de conexión, III: 236, 251
 - de un conjunto de puntos, III: 33
 - de contacto, I: 404
 - de ecuaciones diferenciales, I: 384
 - de grupo, III: 325
 - de grupo simétrico, III: 327
 - de un infinitésimo, I: 151
 - de una matriz cuadrada, III: 59
 - de simetría, III: 312
 - de tangencia, I: 404
 - de transformación, III: 331
- Ordenador, I: 260
- Orientabilidad, III: 245
- Origen, I: 227
- Ortogonal(es), complementos, III: 82
 - funciones, III: 278
 - invariante, I: 289
 - matriz, III: 397
 - proyección, III: 82
 - sistema, II: 351; III: 279, 280
 - transformación, I: 287; III: 101
 - de una forma cuadrática, III: 114
- Ortogonalidad, II: 47
 - de vectores, III: 80, 270
 - propiedad de, II: 345
- Ortonormal, base, III: 80
 - sistema de funciones, III: 281
- Oscilación(es), armónica, III: 292
 - característica, II: 47; III: 293
 - ecuaciones de, II: 35
 - forzada, I: 394
 - libre, III: 292
 - modo de, II: 51
- Ostrogradski, M. V., I: 74, 99, n. 26 p. 169, 179, 203, 205; II: 58, 159
 - fórmula de, I: 203; II: 21, 163
 - Hamilton, principio de, II: 159
- Papaleksi, N. D., I: 424
- Parábola, I: 248
- Parábola de orden n , I: 364
- Paraboloides, I: 273
 - elíptico, I: 269
 - hiperbólico, I: 267, 270
- Paralelas, líneas, III: 133-135
- Paralelas, postulado de, III: 124
- Paralelismo, III: 133
- Parámetro, I: 311
- Paridad, III: 328
- Parte principal, II: 220
- Parseval, M., III: 53
- Parseval, igualdad de, III: 53
- Pascal, B., II: 376
- Perforadora reproductora, II: 378
- Permutación, impar, III: 327
- Perspectiva plana, teorema fundamental de; véase Teorema de unicidad de la teoría de transformaciones proyectivas
- Peterson, K. M., II: 128, 134
 - teorema de, II: 79, 128
- Petrovski, I. G., I: 312; II: 307
- Picard, E., II: 217
- Picard, método de; véase Método de aproximaciones sucesivas
- Pitágoras, I: 40, 45, 51, 69, 229, 240; III: n. 6 p. 140, 202
- Planck, M. K. E. L., II: 307
- Plano, completado, I: 297
 - cubierto densamente, I: 299
 - ecuaciones de, I: 268
 - de las fases, I: 409
 - n -dimensional, III: 169
 - normal, II: 94
 - osculador, II: 94
 - proyectivo, en geometría, I: 297
 - en topología, III: 239-240
 - de simetría, III: 311
 - tangente, II: 98-101
- Pogorelov, A. V., II: 141
- Poincaré, H., I: 74, 84, 99, 301, 416, 424; III: 131, 151, n. 11 p. 151, 154, 222, 243, 252, 254, 258, 260
- Poisson, S. D., II: 30-32, 43, 52, 54, 58
- Poliedros, II: 140

- en un espacio n -dimensional, III: 175
 Polinomio(s), característico, de una matriz, III: 106
 gráfica de, I: 364
 de interpolación, II: 316, 335
 irreducibilidad de, I: 334
 de Legendre, II: 330; III: 282
 de una matriz, III: 117
 simétrico, I: 328; III: 320
 simétrico elemental, I: 328
 trigonométrico, II: 314
 Polo II: 219
 Poncelet, J. V., III: 160
 Pontriagin, L. S., I: 424; III: 260,
 n. 12 p. 260, 261, 311, 359
 ley de dualidad de, III: 261
 Popov, A. S., I: 21
 Posición falsa, método de, I: 366,
 368
 Postnikov, M. M., III: 261
 Potencial, de capa simple, II: 56
 de fuerzas, II: 55
 de velocidad, II: 27, 187
 de un fluido, II: 26
 Primitiva, función, I: 168
 Primo, gemelo, II: 242
 Principio(s), del argumento, I: 360
 de Arquímedes, III: 31-32
 de continuidad de Cantor, III:
 30
 de Dedekind, III: 29
 Probabilidad, II: 269, 273
 adición de, II: 273
 densidad de, II: 306
 distribución de, II: 286
 Proceso aleatorio, II: 299, 306
 Proclo, III: 125
 Producto, III: 26
 de caminos, III: 361
 de ciclos, III: 328
 derivada de, I: 133
 libre, III: 364
 de matrices, III: 60
 no conmutativa, III: 60
 de transformaciones, III: 321
 Prolongación analítica de una función en un dominio complejo,
 II: 170-171, 222-225
 Propagación de ondas, II: 40
 Propiedad intrínseca, II: 127
 Proporcionalidad, III: 56
 Proximidad, III: 359
 Proyección perspectiva, I: 293-5
 Proyectiva(s), aplicaciones: teorema
 fundamental, I: 298
 transformación, I: 298; II: 143
 Punto(s)
 adherente, III: 262-264
 aislado, III: 35
 conjunto de, III: 32
 conjunto lineal de, III: 32
 coordenadas de, I: 227
 de discontinuidad, I: 411
 de ensilladura, I: 413
 de estancamiento, II: 193
 fijo, III: 257
 frontera, III: 35
 impropio, III: 239
 interior de un conjunto, III: 35
 de interpolación, II: 320
 de intersección, I: 230
 de inflexión, I: 148
 límite, III: 35
 ordinario, I: 410
 plano, II: 103
 simétrico, III: 311
 singular, III: 253, 256
 aislado, I: 410
 de tangencia, II: 99

R

- Racionales, números, III: 26
 Radical, III: 394
 de un anillo, III: 404
 Radio de convergencia, I: 220; II:
 173; III: 160
 Radio vector, II: 130
 Rafael, I: 91
 Raíces, cálculo de, I: 364
 múltiples, I: 353
 multiplicidad de, I: 353
 reales, I: 327
 Rama uniforme, II: 182, 227
 Ramanujan, S., II: 240
 Ramificación, punto de, de función
 compleja, II: 182, 226
 Rango, de álgebra, III: 387
 de una matriz, III: 95

- de números hipercomplejos, III: 376
 - de un sistema de vectores, II: 75
 - Rasevski, P. K., III: 215
 - Real, espacio, III: 215
 - número, I: 47; III: 25, 28
 - parte de un cuaternio, III: 381, 382-383
 - punto, I: 298
 - raíz, I: 327
 - recta, I: 298
 - Reciprocidad, ley de, III: 289
 - Recta, línea, II: 123
 - Recta numérica, III: 16, 29
 - Redes de curvas, II: 79
 - Reducción, a una forma canónica, I: 253
 - de una forma cuadrática, III: 111
 - representación, III: 370
 - Reducible, ecuación, I: 292
 - Regular, representación, III: 371
 - sistema de puntos espaciales, III: 342
 - superficies, II: 97
 - Representación completamente reducible, III: 370
 - Representación fiel, III: 391
 - Representación(es), de álgebras asociativas, III: 391
 - equivalentes, III: 367
 - de grado n , III: 368
 - de grupo, III: 367
 - irreducible, III: 370
 - isomorfa, III: 391
 - Reproductora, II: 378
 - Resolubilidad, de grupo, III: 334
 - de problemas con regla y compás, I: 335
 - mediante radicales, III: 346
 - Resoluble, grupo, III: 334
 - mediante radicales, I: 337
 - Reticulo, III: 404
 - modular, III: 406
 - «Resto», término, I: 159
 - Riccati, J. F., I: 395
 - ecuación de, I: 395
 - Riemann, B., I: 80, 81, 99, 121, 312; II: 180, 183, 187, 194, 198, 201-207, 218, 227-230, 240, 246; III: 16, 45, 47, 49, n. 7 p. 144, 197-214, 216, 221, 222, 226, 258, 357
 - espacio de, III: 199
 - geometría de, III: 199
 - hipótesis de, II: 246
 - integral de, I: 165; III: 47
 - métrica de, III: 205
 - superficie de, II: 201-207; III: 258
 - Riesz, F., I: 84
 - Ritz, W., II: 73-74, 359
 - Rolle, M., I: 354, 356
 - teorema de, I: 354
 - Rotación, del espacio, III: 318
 - de una transformación, II: 197
 - Rotaciones, grupo de, III: 326
 - Rotacional, flujo, II: 26
- S
- Saccheri, G. G., III: 126, n. 2 p. 126, 128, 131
 - Safarevich, I. R., I: 338; III: 346
 - Schauder, J., III: 265
 - Schlaefli, L., III: 176
 - Schrödinger, E., II: 73
 - Schwartz, desigualdad de; véase Desigualdad de Cauchy-Bunjakovski
 - Schweikart, III: 127, 131
 - Sección normal, II: 102
 - Segmento parabólico, I: 404
 - Selberg, A., II: 246
 - Semiintervalo, III: 33, 39
 - Semiinvariante, I: 290
 - Semirretículo, III: 405
 - inferior, III: 405
 - superior, III: 405
 - Separación de variables, II: 44
 - Series, I: 206
 - absolutamente convergente, I: 211
 - armónicas, I: 210
 - divergentes, I: 207
 - no uniformemente convergentes, I: 213
 - de potencias, I: 217; II: 169
 - sumables, I: 207, n. 37 p. 207
 - de Taylor, I: 162
 - trigonométricas, III: 283

- uniformemente convergentes, I: 212
- Serre, J. P., III: 261
- Serret, J. A., I: 318
- Seudoesfera, III: 142
- Símbolos, I: 27
- Simetría, III: 311, 317
 - bilateral, III: 312
 - central, III: 312
 - centro de, III: 312
 - circular dirigida, III: 336
 - eje de, III: 312
 - especular, III: 311
 - de orden n , III: 312
 - respecto a ejes, III: 342
- Simplex, III: 178, n. 22 p. 178
 - orientado, III: 250
- Simpson, T., II: 324-330, 360, 368, 369
 - fórmula de, II: 325
- Singular, punto, I: 410; III: 256
- Sistema físico, oscilaciones características, I: 391; III: 293
- Sistema, algebraico, III: 309, 406-408
 - binario, III: 404
 - de ecuaciones lineales, III: 83
 - homogéneo de ecuaciones lineales, III: 96; *véase también* Ecuaciones de primer grado de intervalos, III: 29
 - normal, II: 351
 - ortogonal de funciones, III: 282, 285
 - normal, II: 351
 - posicional, I: 29
- Sitnikov, K. A., n. 12 p. 260, 261
- Smidt, O. Ju., III: 311, 373
- Smirnov, Ju. M., III: 266
- Smirnov, V. I., I: n. 33 p. 194
- Snirelman, L. G., II: 138; III: 260
- Sobolev, S. L., II: 69, 73
- Solución(es), admisible, II: 68, 69
 - algebraica, I: 320
 - de ecuaciones diferenciales, I: 384
 - generalizada, II: 68, 69
 - unicidad de, I: 401; II: 29
- Stepanov, V. V., I: 425; II: 54, 302; III: 255
- Stevin, S., I: 64
- Sturm, S., I: 318, 358
 - teorema de, 357
- Subálgebra, III: 393
- Subanillo, III: 399
- Subespacio, III: 74
- Subespacio generado por un sistema de vectores, III: 74-75
- Subgrupo(s), III: 330-334
 - generadores de, III: 330
 - invariante, III: 352
 - normal, III: 351
 - normales no triviales, III: 358
 - simple, III: 358
- Subnikov, A. V., III: 343
- Sucesión, convergente en casi todo punto, III: 46
 - fundamental, III: 30
 - minimizante, II: 16
 - de números reales, III: 30-31
- Suceso(s), aleatorio, II: 271, 296, 303
 - cierto, II: 269
 - estocástico, II: 271
 - imposible, II: 269
 - independiente, II: 275
 - intersección de, II: 274
 - mutuamente excluyentes, II: 272
 - necesario, II: 269
 - unión, II: 272
- Suma, de conjuntos, III: 19
 - derivada de una, I: 132
 - de Fejér, II: 349
 - parcial, I: 207
 - de series, I: 207
 - trigonométrica, II: 247, 248
 - de vectores, I: 262
- Sup., I: 213, n. 39 p. 213
- Superficie(s),
 - analítica, II: 125
 - cerrada(s), III: 237, 245, 246
 - clase de, II: 139
 - convexa, II: 140-141
 - orientable, III: 250-251
 - con curvatura variable de Gauss, II: 121
 - conexión de, III: 236
 - congruente, II: 128
 - de curvatura media nula, II: 111
 - de dos caras, III: 236
 - de una sola cara, III: 236
 - deformabilidad de, II: 142

deformación de, II: 115, 125, 142
 desarrollable, II: 120
 de discontinuidad, II: 68
 ecuaciones de, I: 266
 familia de, II: 134
 género de, II: 230; III: 236
 geodésica, III: 210
 lisa, II: 99
 mínima, II: 111
 de Möbius, III: 236-237
 no cerrada convexa, II: 142
 no lisa, II: 140
 no regular, II: 137
 normal, III: 236
 regular, II: 97
 de revolución de menor área, II: 147
 teoría de, II: 96
 Superposición, método de, II: 44
 Suslin, M. Ja., III: 41
 Sustracción, III: 26
 Sylvester, J. J., I: 358; II: 245

T

Tablas, II: 273
 de dos entradas, II: 374
 de una entrada, II: 374
 Tabulador, II: 378
 Tales, I: 40
 Tangente, I: 122; II: 86, 87
 Tangentes, campo de, I: 399
 Tartaglia, N., I: 63, 318, 321, 322
 Taurinus, III: 127, 131
 Taylor, B., II: 320, 321
 fórmula de, I: 157, 191
 serie de, I: 162
 Temperatura, gradiente de, II: 28
 Tensor, II: 16
 Teorema,
 de Abel, I: 332; II: 171
 de Bolzano-Weierstrass, I: 347; III: 36
 de Bonnet, II: 128
 de Budan, I: 357
 de Cauchy, II: 210
 central del límite, II: 290
 de Codazzi, II: 80, 128
 de contracción, II: 364

de dualidad, III: 260
 de Euler, II: 103, 105
 fundamental, del álgebra, I: 338-352; II: 169
 de la perspectiva plana, I: 295-297
 de Gauss; véase Ostrogradski, fórmula de
 de Green, II: 163
 de Hopf, III: 256
 del límite central, II: 290
 de Meusnier, II: 104
 de Peterson, II: 79, 128
 de Picard, II: 217
 del punto fijo, de Brouwer, III: 198, 258
 de Rolle, I: 354
 de Sturm, I: 357
 de Sukowski, II: 190, 193
 de unicidad, para funciones analíticas, II: 222
 de la teoría de transformaciones proyectivas, I: 294
 del valor medio, I: 154
 de Wedderburn, III: 395
 de Weierstrass, II: 337
 de Wilson, II: 265
 Teoría, algebraica de números, II: 240
 analítica de números, II: 240
 cualitativa de ecuaciones diferenciales, I: 415; III: 253-254
 de curvas, II: 80
 elemental de números, II: 240
 de funciones de variable real, III: 16-17
 de Galois, I: 332
 geométrica, de funciones de números, II: 240
 de Newton, de diámetros, I: 237
 de números, II: 235
 de operadores, III: 297
 Tihonov, A. N., III: 266
 Tolomeo, C., I: 58, 60, 63
 Topológico(a), aplicación, III: 263
 espacio, III: 193, 262, 265
 estructura, III: 261
 grupo, III: 358
 imagen, III: n. 10 p. 259
 invariante, III: 247, 259, 362
 producto, III: 242

- propiedad, III: 232
 - transformación, III: 232
 - Topología, III: 231
 - ciclos en, III: 249
 - combinatoria, III: 260
 - método de, III: 243-244
 - conjuntista, III: 259
 - n -dimensional, III: 259
 - orientaciones compatibles en, III: 245
 - de superficies cerradas, III: 258
 - Toro, III: 241
 - de una cara, III: 238
 - Torricelli, I: 97
 - Torsión, II: 95; III: 250
 - Tractriz, III: 142
 - Trascendente, función, II: 217
 - número, II: 240
 - Transformación(es), afín, I: 280; III: 162
 - bicontinua, III: 362
 - conforme, II: 197, 200
 - conjugada, III: 323, 324, 332, 333
 - de coordenadas, I: 255; III: 102
 - cuasi-conformes, II: 204
 - degenerada, III: 101
 - general afín, I: 280
 - grupo de, III: 351
 - de grupo, III: 163
 - grupos continuos de, III: 336
 - homogénea, III: 99
 - identidad, III: 330
 - inversa, III: 99
 - lineal, III: 98
 - de Lorentz, I: 301, 303
 - matriz de los coeficientes de, III: 100
 - no degenerada, III: 101
 - orden de, III: 331
 - ortogonal, I: 287; III: 101
 - periódica, III: 331
 - producto de, III: 321
 - proyectiva, I: 298; II: 143; III: 158
 - simétrica, III: 314
 - singular, III: 101
 - no singular, III: 101
 - topológica, III: 232
 - unitaria, III: 358
 - uno a uno, III: 315
 - Transposición, III: 328
 - Traslación, III: 318
 - Traspuesta, III: 64
 - matriz, III: 102
 - Traza, III: 372
 - Treffitz, J., II: 73
 - Triangulación, III: 244
 - tridimensional, III: 247
 - Triangular, desigualdad, III: 196
 - ley, III: 264
 - Tschirnhausen, E. W., I: 318, 326
- U
- Ulug Begh, I: 62
 - Unicidad, de descomposición, III: 401
 - propiedades de, II: 222
 - de solución, I: 401; II: 29
 - Unicidad, teorema de, para funciones analíticas, II: 222
 - de la teoría de transformaciones proyectivas, I: 295
 - Unidad, III: 401
 - Uryson, P. S., III: 259, 260, n. 11 p. 260, 263, 266
- V
- Vagner, V. V., III: 215
 - Vallée-Poussin, C. J., de la, II: 314
 - Van der Waerden, B. L., I: 320
 - Variable, I: 65
 - aleatoria, II: 286
 - cambio de, I: 175
 - continua, I: 54
 - dependiente, I: 103
 - independiente, I: 103
 - separación de, II: 44
 - Variable, magnitud, I: 65, 110
 - Variación, II: 154
 - Variacional, problema, II: 157
 - Varianza, II: 287
 - Varichak, III: 151
 - Variedades, III: 241-243
 - aplicación continua de, III: 257

- conexión de, III: 251
 - irreducible, III: 403
 - lisas, III: 255
 - multi-dimensionales, III: 259
 - Vavilov, S. I., I: n. 1 p. 92
 - Veblen, O., III: 260
 - Vector(es), I: 261; III: 65-66
 - combinación lineal de, III: 70
 - de un espacio infinito-dimensional, III: 275
 - flujo, II: 20
 - inverso de un, I: 262
 - longitud de, III: 80
 - n -dimensional, III: 270
 - dos a dos, III: 272
 - velocidad, II: 184
 - Vectorial(es), álgebra, I: 261; III: 381
 - campo, II: 15; III: 252
 - continuo, III: 252
 - coordenadas, I: 263; III: 73
 - función, II: 130
 - operaciones, III: 65
 - parte de un cuaterno, III: 380, 382
 - suma, I: 262
 - de dos subespacios, III: 75
 - Velocidad potencial, II: 27, 187
 - Vieta, F., I: 61, 63, 317, 319, 326, 328; III: 344
 - Vinogradov, I. M., I: 87; II: 240, 241, 246, 256-264, 316; III: 402
 - Vonoroï, G. F., II: 241
- W
- Wachter, III: 127
 - Wallace, A. R., III: 131
 - Wallis, J., III: 125
 - Wedderburn, III: 394, 395, 404
 - teorema de, III: 395
 - Weierstrass, K. T., I: n. 14 p. 46, 83, 99, 129; II: 314, 337; III: 30, 31, 36
 - principio de, III: 30
 - teorema de, II: 337; III: 36
 - Weyl, H., III: 398
 - Wilczynski, E. J., II: 143
 - Wilson, teorema de, II: 265
- Z
- Zenón, I: 46
 - Zolotarev, E. I., I: 318; II: 240, 314; III: 402
 - Zukovski, N. E., I: 344; II: 190, 191, 193
 - teorema de, II: 190, 193